INVESTIGATING COMPLEXITY IN TRANSCRIPTOME EXPRESSION, REGULATION, AND EVOLUTION USING MATHEMATICAL MODELING

By

Nicholas Louis Panchy

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Genetics – Doctor of Philosophy

2017

ABSTRACT

INVESTIGATING COMPLEXITY IN TRANSCRIPTOME EXPRESSION, REGULATION, AND EVOLUTION USING MATHEMATICAL MODELING

By

Nicholas Louis Panchy

To date, gene expression has been characterized in over one thousand species across more than a million experimental conditions. With this wealth of data, it is possible to investigate the role that differential expression has in key biological processes, such as development, stress response, and cell division. However, the complexity of the transcriptome makes the analysis of expression challenging, as a single genome can contain thousands of genes as well as millions of potential regulatory interactions shaped by more than a billion years of evolution. To address this complexity, we can use the language of mathematics to create models of gene expression, regulation, and evolution that define the system in a testable format. In the following chapters, I will present research that applies mathematical modeling to the identification and regulation of cyclically expressed genes as well as the evolution of transcriptional regulators following whole genome duplication.

Cyclically expressed genes were studied in two systems. First, I investigated day-night cycling or 'diel' genes in *Chlamydomonas reinhardtii*. Diel genes were identified *de novo* using two models of cyclic expression that jointly classified half of all genes in *C. reinhardtii* as diel expressed. To understand the regulation of diel expression, I clustered diel genes according their peak of expression, or 'phase', and searched for *cis*-regulatory elements enriched (CREs) in the promoters of each cluster. While I found putative CREs corresponding to each cluster, using these CREs to predict diel expression using machine learning performed poorly compared to previous models of expression regulation. Therefore, I changed systems to *Saccharomyces*

cerevisiae and studied cyclic expression during the cell cycle. Here, I applied machine learning models to predict cell-cycle expression using regulatory interactions from four different data sets. These models out performed the previous model of cyclic expression when using regulatory interactions defined by chromatin-immunoprecipitation, transcription factor knockout experiments, and position weight matrices. Further gains in performance were obtained by combing interactions across data sets and using co-regulation by pairs of regulators involved in feed-forward loops. The most important interactions for predicting cell-cycle expression included not only known cell-cycle regulators but also two groups of transcription factors not previously identified as being involved in cell-cycle regulation.

The evolution of transcriptional regulation was studied in *Arabidopsis thaliana,* which has undergone several rounds of whole genome duplication (WGD), after which transcriptions factors (TFs) are preferentially retained. Here, I applied maximum likelihood estimation to infer the most likely ancestral expression and regulatory state of pairs of duplicate TFs prior to WGD. Comparing this ancestral state to the existing TF duplicates, I found that one duplicate, the "ancestral' copy, tends to retain the majority of ancestral expression state and CREs, while the other 'non-ancestral' copy loses ancestral expression and CREs, but also gains novel CREs instead. Modeling the evolution of TFs pairs using as system of ordinary differential equations, I demonstrated that the partitioning of ancestral states amongst duplicates is not random, but occurs because the loss of ancestral expression occurs orders of magnitude faster in the first copy than in the second. This suggests that TFs duplicate pairs are preferentially maintained such that one copy is 'ancestral' and the other is not. Taken as a whole, the research in this dissertation demonstrates how mathematical modeling can be applied to studying the expression, regulation and evolution of the transcriptome.

Dedicated to…
…my family at home
…my family in the Shiu Lab
…my family from Providence

Thank you all for your love, support, and patience

ACKNOWLEDGEMENTS

First, I would like to thank Dr. Shin-Han Shiu for agreeing to be my mentor. It was talking with Shin-Han that helped convince me that MSU was where I wanted to do my graduate studies and over the years his guidance, advice, and critique have been invaluable to my growth as a scientist.

I would also like to thank my committee: Drs. David Arnosti, Eva Farre, and Chris Adami for their input and feedback on my research. I also greatly appreciate the candor with which they were willing to discuss the realities of being an academic as it stopped me from becoming discouraged during my time as a student. Additionally, I would like to thank the Genetics Graduate Program- particularly Barb Sears, Brian Schutte, Cathy Ernst, and Jeannine Lee- as well as all the Genetics students I have shared my time with for helping navigate life as a graduate student.

I owe a great deal to the past and current members of the Shiu Lab- Melissa, Gaurav, Gunagxi, Sahra, Alex, Dave, Ming, Zing, Johnny, Beth, Christina, Peipei, Siobhan, Reid, and the many visiting scholars and undergraduate students we worked with over the years. After all the time we have spent together inside and outside of the lab, I feel that share a special sense of camaraderie with you all. I am looking forward to hearing of your future successes even though I may no longer be with you.

Finally, I would like to thank my parents for encouraging me to pursue my education and supporting me through my years in college and graduate school and my family at Providence PCA for providing me a home in Michigan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

ANOVA	Analysis of variance
Asy	Asymmetry score
AUC-ROC	Area under the curve of the receiver operating characteristic
BLAST	Basic alignment search tool
bp	base pair
ChIP	Chromatin immunoprecipitation
CPM	Counts per million
CRE	*Cis*-regulatory element
Ctrl	Control
DAP-Seq	DNA affinity purification sequencing
Diel	24-hour, day night period
DNA	Deoxyribonucleic acid
DFT	Discrete Fourier transform
Diff	Differential Expression
FFL	Feed –forward loop
FPKM	Fragments per kilo-base of transcript per million mapped reads
G1	Initial growth (cell-cycle)
G2	Intermediate growth (cell-cycle)
GEO	Gene Expression Omnibus
GO	Gene ontology
GRN	Gene regulatory network

K_a	Non-synonymous substitution rate
K_s	Synonymous substitution rate
Kb	kilo-basepair
IQR	Inter-quartile range
LightDev	Light and Development
miRNA	micro RNA
M	Cell division (cell-cycle)
MAD	Mean absolute deviation
MSU	Michigan State University
NCBI	National Center for Biotechnology Information
ODE	Ordinary differential equation
PBM	Protein binding microarray
PCC	Pearson's correlation coefficient
PCR	Polymerase chain reaction
pCRE	putative CRE
PHYLIP	Phylogeny Inference Package
PWM	Position weight matrix
pv	P-value
r^2	Coefficient of determination
RAxML	Randomized Axelerated Maximum Likelihood
RIN	RNA Integrity Number
RMA	Robust Multichip Average
RNA	Ribonucleic acid

RNA-Seq	RNA sequencing
RPKM	Reads per kilo-base of transcript per million mapped reads
S	DNA replication (cell-cycle)
SRA	Sequence Read Archive
SVM	Support Vector Machine
TAP	Tris-Acetate-Phosphate
TF	Transcription factor
TPM	Transcripts per million
WEKA	Waikato Environment for Knowledge Analysis
WGD	Whole genome duplication
ZT	Zeitgeber time

CHAPTER 1: INTRODUCTION

Following the advent of microarray (Schulze and Downward, 2001; Hoheisel, 2006) and high-throughput sequencing (Reuter et al., 2015; Goodwin et al., 2016) technology, gene expression has been inferred using transcript quantification in over 3300 species, with more than 400 having in excess of 100 samples publically available through online databases (GEO, https://www.ncbi.nlm.nih.gov/geo/summary/?type=taxfull). With such a breadth of expression data available, in terms of transcriptome coverage, organisms, and conditions, it has become possible to characterize genes using their expression profiles. The analysis of these profiles has been applied to a variety of research questions: the progression and outcome human disease (Henriksen and Kotelevtsev, 2002; van 't Veer et al., 2002; Bergholdt et al., 2009; Cooper-Knock et al., 2012), understanding the basis for variation in quantitative phenotypes (Jiménez-Gómez et al., 2010; Jimenez-Gomez et al., 2011; Nica and Dermitzakis, 2013; Albert and Kruglyak, 2015), and predicting the phenotype/functions of genes (van Noort et al., 2003; Takagi et al., 2014; Lloyd et al., 2015; Uygun et al., 2016). Yet, while expression profile analysis can be useful for identifying and classifying genes, the question remains as to how patterns of expression are established and maintained. One approach to understanding how expression patterns are regulated is the use of mathematical modeling: the representation of a system using mathematical objects (variables, operators, equations, etc). For gene expression in particular, this involves defining using set of explanatory variables to predict the expression of genes as accurately as possible in order to answer a biological question about the genes, their regulators, or the dynamics of the system. Although the molecular mechanisms that regulate gene expression are understood (Lee and Young, 2000; Lelli et al., 2012; Voss and Hager, 2014) and broad patterns of expression can be inferred from sequence alone (Beer and Tavazoie, 2004), modeling gene expression remains a challenging task, particularly in response to specific

environmental conditions (Zou et al., 2011), cellular location (Uygun et al., 2017), and time (Panchy et al., 2014). My research focuses on the application of differential equations and machine learning models to understanding the regulation of cyclic expression and evolution of regulatory systems, but many different modeling approaches have been applied to an equally varied set of biological questions.

MOLECULAR MECHANISMS OF GENE REGULATION

As gene expression is often quantified by the level of mRNA transcript, approaches to modeling gene expression are guided by what is known about the regulation of transcription at the molecular level. The transcription of a gene is primarily (but not exclusively) regulated at the initiation stage when the RNA-polymerase complex is recruited to the promoter region upstream of the transcription start site (Lee and Young, 2000). This promoter region contains a core promoter element, which is ubiquitous in function across eukaryotic genes, that binds the components of the RNA-polymerase complex, which is also common across eukaryotes. However, the core promoter alone is insufficient for regulation of transcription *in vivo*, and additional factors, called transcription factors (TFs), are required to enhance or repress RNA-polymerase binding and activity (Lee and Young, 2000). Because modeling is primarily focused on differences in expression either between genes or in a single gene across time or a set of conditions, the common core elements can be ignored in favor of the activity of TFs.

The affinity of TFs for a particular promoter primarily depends on regions of DNA known as *cis*-regulatory elements (CREs), such that the presence or absence of these elements represent TF regulation. However, there is not a 1-to-1 relationship between TFs and CREs, but rather a single TF can bind multiple CREs with varying degrees of sequence similarity, and a single base change may or may not disrupt the binding potential of an element depending on the

position of the change (Badis et al., 2009). Furthermore, whether or not a TF interacts with a promoter also depends on chromatin state, nucleosome positioning, histone modification, and cooperativity with other TFs (Lee and Young, 2000; Lelli et al., 2012; Voss and Hager, 2014). Although the presence or absence of CREs is relatively fixed in a given genome, the accessibility of chromatin, histone-code and concentration of other TFs are all dynamic, meaning that CREs alone are often insufficient to determine if a TF regulates a gene under a specific set of conditions and thus plays a role in regulation. Therefore, the first step in modeling gene expression is determining how to identify a set of relevant regulatory interactions.

IDENTIFYING REGULATORY INTERACTIONS

The types of evidence that can be used to identify regulatory interactions can be divided into three broad categories: directly assaying protein-DNA interactions, prediction of TF binding from promoter sequence, and inferring interaction between genes based on expression variation. The DNA sequence(s) that a protein will bind to can be assayed either *in vitro* using protein binding microarrays (Bulyk, 2007; Berger and Bulyk, 2009), or *in vivo* with chromatin immunoprecipitation (Buck and Lieb, 2004; Furey, 2012) or DNA affinity purification (O'Malley et al., 2016). Because all approaches used to define regulation are based on binding sequences, it is necessary for a genome to be sequenced and annotated so that the recovered sequence can be mapped to the promoter region of potential target genes. Often, post-processing is required to address noisy TF binding (i.e. false-positive interactions). For high-throughput sequencing approaches in particular (e.g. ChIP-seq and DAP-seq), high-confidence binding sites can be identified by mapping sequences reads to the genome and using software (Zhang et al., 2008) to call "peaks" where multiple-reads overlap the same sequence (Feng et al., 2012; Landt et al., 2012; O'Malley et al., 2016). However, experimental evidence may not be available for

every TF in an organism and, given that there are dozens of sequenced eukaryotic genomes with more than 1000 TF genes (Kummerfeld and Teichmann, 2006), assaying binding of the entire set of transcription factors may not be feasible in some cases.

In the absence of direct binding information for TFs, the presence of "putative" *cis*-regulatory elements (Beer and Tavazoie, 2004; Zou et al., 2011; Liu et al., 2015; Uygun et al., 2017) may be inferred from the promoter sequence of target genes (reviewed in Das and Dai, 2007 and Li et al., 2015). By using computational approaches for binding site prediction (software such as AlignAce, YMF, MEME), it is possible to identify pCREs using promoter sequences alone (Hughes et al., 2000; Sinha and Tompa, 2003; Bailey et al., 2006; Bailey et al., 2009). There are also machine learning and deep learning methods for integrating multiple types of omics data, including DNA accessibility, chromatin structure, histone marks, and available binding site information from assays like ChIP (Pique-Regi et al., 2011; Hoffman et al., 2012; Alipanahi et al., 2015; Li et al., 2016). Furthermore, transcription initiation events associated with *cis*-regulatory elements can produce non-coding RNAs that can be captured by Global Run-On sequencing and used to infer pCREs from data (Danko et al., 2015). Predictive methods do not necessarily connect pCREs to the TFs that bind them, but pCREs derived from the promoters of co-expressed genes do show similarity to known TF binding motifs (Uygun et al., 2017), which suggests that presence of pCREs do in fact reflect the binding potential of TFs. However, whether assayed or predicted, TF-target gene interactions identified based on TF binding suffer from the same drawback: binding potential does not guarantee regulatory function as actual TF binding can be greatly affected by small changes in sequence (Kwasnieski et al., 2012) and the presence of other TFs at the promoter site (Spivak and Stormo, 2016).

Alternatively, the interaction between TFs and target genes can be inferred based on changes in gene expression. In this case, the presence of an interaction is assumed if two genes share a "coordinated" pattern of expression, though what constitutes "coordinated" varies with approach. Coordination of expression has been characterized using mutual information (Margolin et al., 2006; Faith et al., 2007), regression (Geeven et al., 2012), differential equations (Honkela et al., 2010), and Bayesian networks (Friedman et al., 2000). However, the results of the Dialogue on Reverse Engineering Assessment and Methods (DREAM) network inference challenge, an open challenge to infer gene regulatory networks from a standard set of synthetic and actual expression data (Marbach et al., 2010; Marbach et al., 2012), suggest that ensemble methods that combine multiple approaches have the best performance when predicting an artificially generated gene regulatory network with 195 regulators and 1643 genes. However, even the best ensemble methods perform poorly when applied to *Escherichia coli* (296 regulators, 4297 genes) and no better than random guessing on *Saccharomyces cerevisiae* (183 regulators, 5677 genes), suggesting that the performance of expression based methods decline as the size of the network increases (Marbach et al., 2012). Furthermore, methods that predict interactions based on expression tend to exhibit common errors, such as inferring relationships between co-regulated genes where none exist (fan-out error), inflating the number of interactions possessed by highly connected target genes (fan-in error), and inferring "shortcuts" between the beginning and end of pathways (cascade error)(Marbach et al., 2010). Including expression from TF-knockout experiments helps reduce fan-in and fan-out error, and TF-knockout data has been used to directly infer interactions (Reimand et al., 2010). However, TF-knockout data was not useful for addressing cascade errors in a model context (Marbach et al., 2012). In *S. cerevisiae*, most TF-target interactions derived from TF-knockout studies lacked evidence of direct, *in vivo*

binding when compared with ChIP-Chip binding data, though there was often was evidence of interaction through an intermediate TF. Keeping these caveats in mind, interactions inferred from expression data can provide useful information for modeling expression, and results presented later suggest that combing both binding data and TF-knockouts improves predictions of expression. Yet interactions are only half of the equation: before a mathematical model of expression can be made, what the model is trying to predict must also be defined.

DEFINING GENE EXPRESSION

Compared to identifying regulatory interactions, defining what the model is trying to predict may seem trivial as the question often comes before the model. Yet, even if the set of target genes and the pattern of interest are known beforehand, it is still be necessary to decide how to define gene expression. At its most basic, modeling expression can be treated as a quantification problem or a classification problem. Using stress response as an example, quantification would involve comparing two continuous expression values before and after stress, while classification would involve categorizing genes as up-regulated, down-regulated or unchanged following stress. There is no "best" way to treat expression in this regard, but rather how expression is defined should be guided by the question at hand; is it important to know that gene expression changes or the magnitude of the expression changes that occur? Making this distinction is an important first step to determining what type of data is needed, how to treat that data, and what modeling approach to use.

Even if expression is treated as a classification problem, categorizing or identifying expression patterns often begins with quantifying the amount of mRNA transcript in a sample. Several technologies are currently available to quantify transcript levels, including Northern blotting (Fernyhough, 2001), fluorescence in situ hybridization (Femino et al., 1998), reverse

transcription PCR (Bustin, 2000; Nolan et al., 2006), microarrays (Xiang and Chen, 2000) and high-throughput sequencing (Wang et al., 2009). Of the more than two million expression data sets publically available through GEO, 96.6% are derived from microarray or sequencing, with PCR a distant third (1.3%). Microarray data from GEO can be accessed and analyzed using BioConductor (Gentleman et al., 2004; Davis and Meltzer, 2007) while sequencing data needs to pre-processed, mapped, and quantified (reviewed in Conesa et al., 2016). Metrics for quantifying sequenced reads comes in two types: (1) counts/transcripts per million (CPM/TPM), in which the number of reads/assembled transcripts is adjusted based on the total number of mapped reads/transcripts in millions, and (2) reads/fragments per kilobase of transcripts per million mapped reads (RPKM/FPKM). In general, RPKM/FPKM is preferred for comparing expression within samples because the longer transcripts tend to produce more reads, while CPM/TPM preferred for comparing across samples/species (Conesa et al., 2016).

Expression can also be quantified as the difference in expression between genes across samples (e.g. treatment vs. control). For microarray data, BioConductor provides a protocol for differential expression (see https:// www.bioconductor.org/help/workflows/arrays/), but choosing the best approach for sequencing data depends on the number of experimental conditions, number of replicates per condition, sample size and available computational resources (see Soneson and Delorenzi, 2013 and Rapaport et al., 2013). Differential expression can also be applied to classification problems. In this case, the significance and direction of differential change can be used to classify genes as up-regulated, down-regulated, or not changed under specific treatments, though it is not uncommon to require a minimum level of change relative to control conditions as well (Kilian et al., 2007; Wu et al., 2015; Uygun et al., 2017). In the case of multiple treatment conditions, this scheme can be applied to each condition independently, but

the number of possible classes will increase quickly (3^N) and require multiple-hypothesis testing. Software like edgeR (Robinson et al., 2010) and DESeq (Anders and Huber, 2010) can be used to directly test the significance of defined patterns of differential and non-differential expression across multiple conditions; however, if specific patterns of expression are not known, clustering can be applied to identify patterns of expression de-novo (Kerr et al., 2008; Oyelade et al., 2016).

Differential expression is not the only criteria for classifying genes by expression. In the case of long time series, progressive or repeated change in expression relative to the mean or starting level of expression may be of interest. A good example of this is cyclic patterns of gene expression, such as occurs across the cell-cycle (Spellman et al., 1998) or in response to the circadian rhythm (Chen et al., 1998; Sukumaran et al., 2010). Approaches to identifying cyclic expression have employed both models of cyclic expression (Straume, 2004; Hughes et al., 2010) and the underlying periodicity expected of cyclic expression (Wichert et al., 2004; Panchy et al., 2014). Although these models are specific to cyclic expression, the same sort of approach can be applied to any pattern of expression.

A final consideration for classifying genes by expression is how to define a negative set, i.e. a set of genes without the desired pattern. Often, this is not as simple as using all other genes because genes which lack the target pattern of expression are not all alike. Therefore, it can be advantageous to define a negative set of genes using its own, separate pattern of expression. For example, when classifying salt-responsive genes, Zou et al. used negative genes that were not differentially expressed under any stress because of possible cross-talk between genes expressed under different stress conditions (Zou et al., 2011). The decision of how to define a negative set will also be influenced by what approach is used to model expression because certain methods, such as machine learning, are more sensitive to the choice of negative examples. Ideally, the

overall process of defining expression will be co-simultaneous with model development in order to avoid conflict.

APPROACHES FOR MODELING GENE EXPRESSION

Defining the modeling problem will also influence the approach used to model gene expression. At its most basic, the expression of a gene can be discretized as either active (1) or inactive (0), allowing interactions between genes and their regulators to be defined using logical operations (i.e. AND, OR, NOT) (Karlebach and Shamir, 2008; Ay and Arnosti, 2011). In this form, expression can be modeled using Boolean networks (Glass and Kauffman, 1973; Thomas, 1973; Kauffman et al., 2003). Boolean networks have been used to study the robustness and stability of GRNs in a variety of systems including *S. cerevisiae (Kauffman et al., 2003)*, human cancers (Shmulevich et al., 2003; Trairatphisan et al., 2016), and *Drosophila melanogaster* (Sánchez and Thieffry, 2001; Yuh et al., 2001; Albert and Othmer, 2003). This method can also be extended to cases where there is imperfect information about regulatory interactions by using probabilistic Boolean networks (Shmulevich et al., 2002; Shmulevich et al., 2003). However, Boolean networks fail to accurately capture the behavior of certain biological interactions, particularly cases where a gene negatively regulates its own expression (Karlebach and Shamir, 2008; Ay and Arnosti, 2011).

Another issue is that analyzing Boolean networks becomes increasingly difficult with increasing size of the GRN being studied (Karlebach and Shamir, 2008). The number of possible global states in a network grows according to the number of states per gene (k) and the number of genes (n) in exponential fashion (k^n). Therefore, the number of global states in even a relatively small genome such *E. coli* K-12 becomes prohibitively large ($2^{4500} \sim 10^{1350}$). For this reason, it is often beneficial to cluster co-expressed genes together so that, instead of modeling

the global pattern of expression across all genes, the problem becomes correctly assigning genes to a finite number of co-expression modules. This clustering approach was taken by Beer and Tavazoie (Beer and Tavazoie, 2004), who used Bayesian networks to assign *S. cerevisiae* genes to one of 49 representative clusters, and was extended by Yuan et al. (2007) with the use of a naive Bayes classifier. Though Beer and Tavazoie (2004) used k-means clustering in order to construct gene expression modules, other clustering methods are available, including hierarchical clustering, self-organizing maps, self-organizing tree algorithms (Yin et al., 2006), as well as more than a dozen different distance metrics (Jaskowiak et al., 2014).

Alternatively, classification, either in the form of discretization or clustering, can be avoided altogether and the quantitative measures of expression taken from experimental data can be modeled directly. Using linear models, gene expression can be modeled from only expression data by assuming each regulator functions independently and its net effect on the target gene is summarized by a singular weight value. However, while linear models have been used to infer regulatory interactions (Yeung et al., 2002; Bansal et al., 2006) and understand risk factors in human disease (Li et al., 2014; Trabzuni et al., 2014), they cannot applied to questions about the dynamics of molecular regulation because the behavior of this system is non-linear (Karlebach and Shamir, 2008). In contrast, thermodynamic models (Bintu et al., 2005; Segal et al., 2008) and Michaelis–Menten kinetics (Nachman et al., 2004) have been used to account for the concentration-dependent nature of TF binding to CREs using probabilistic binding and non-linear functions, respectively. Notably, the Michaelis-Menten equation was derived as a solution to a system of ordinary differential equations (ODEs) describing enzyme kinetics under certain assumptions (Schnell, 2014; Wong et al., 2015). Other systems of ODEs have been used to incorporate different assumptions and variables into models of expression such variable cell

mass and volume (Chen et al., 2004; Li et al., 2008), spatial context and diffusion (Eldar et al., 2002; Jaeger et al., 2004), and separate binding mechanics for protein regulators and microRNAs (Zhang et al. 2014; Hong et al. 2015). However, both systems of ODEs and thermodynamic models are sensitive to the choice of regulatory interactions, such that the erroneous omission or addition of a single regulator can potentially have a significant effect on the outcome (Ay and Arnosti, 2011).

Though they are most obvious in complex models of quantitative expression measures, all modeling approaches described so far make assumptions about how regulators function to control the expression of their targets. Alternatively, the problem of modeling gene expression can be approached by trying to "learn" what features are important for regulating expression using machine learning algorithms (reviewed in Libbrecht and Noble, 2015). Rather than create an explicit model of how gene expression is regulated, these approaches employ programs designed to optimize some task (in this case, the prediction of gene expression) from a set of features (regulatory interactions and any other data). This approach represents a double edged sword in that machine learning algorithms can incorporate many different types of data without prior knowledge of how they function in a system and assess their importance to controlling gene expression, but little can be interpreted about why a specific feature is important from the resulting model. Traditional machine learning algorithms, such as support vector machines and random forest, have been applied to understand the effects of combinatorial regulation (Zou et al., 2011), nucleosome positioning (Liu et al., 2015), and tissue-specific regulation (Uygun et al., 2017) on gene expression in *Arabidopsis thaliana* as well as the influence chromatin state on general expression in *Caenorhabditis elegans* (Cheng et al., 2011) and human cell lines (Cheng et al., 2011; Dong et al., 2012). Furthermore, so called "deep learning", which uses multi-layered

neural networks, has recently been applied to predict gene expression using expression data (Chen et al., 2016b) and histone modification (Singh et al., 2016). This method in particular holds great promise for biological research, not only because it has the potential to outperform traditional machine learning methods (Singh et al., 2016), but also because there have recent, rapid advances in this technology (Chen et al., 2016a; Min et al., 2016; Silver et al., 2016; Fernando et al. 2017) that promise new opportunities for applying deep learning to the biological sciences.

APPLICATIONS FOR GENE EXPRESSION MODELS

Ultimately, the objective of all expression models is to accurately predict expression in the target set of genes, and this predictive function alone is sufficient to answer biological questions. The resulting models can also be used to explore the dynamics of the GRN being modeled as well as discover new elements important for regulating expression. An example of a direct application of predictive models includes Li et al. (2008) who built an ODE model of cell division (including the expression of key regulatory and structural genes) in the stalked cells of *Caulobacter crescentus*. The parameters of the model were fit using the expression values in wild-type cells and subsequently validated by testing if known mutant phenotypes mutants could be reproduced by modifying the network to mimic the mutation. Expect when a mutation involved a process outside of the model (e.g. phosphorylation of regulators), the Li model was able to recreate mutant phenotypes and was subsequently used to predict the phenotype of previously uncharacterized mutants. Similarly, Chen et al. (2004) constructed an ODE model of the cell cycle of *S. cerevisiae* that could accurately model the wild-type cell-cycle as well as the phenotypes of 92% of characterized mutants. In some cases, predictions made about novel mutants were independently validated by another research group (Archambault et al., 2003).

However, Chen et al. noted that, though their model robustly predicted expression during the wild-type cell-cycle, accurately predicting mutant phenotypes was more sensitive to small changes in parameter values (2004). Hence, because of this sensitivity, it is reasonable to treat novel phenotype predictions with skepticism even when the underlying model accurately characterizes expression under normal conditions.

Predictions are not the sole purview of expression models, and often it is the model itself which is of interest, as it can be used to explore the dynamics of the system. Li et al. (2004) constructed a Boolean network of 11 key regulators of the *S. cerevisiae* cell-cycle. They found that all possible initial conditions eventually progressed into one of seven steady states, with most (86%) initial conditions resulting in a steady state representative of the G1 phase, the resting state of the cell cycle. Furthermore, artificially inducing the cell-cycle (i.e. activating Cln3) in the model resulted in an unstable G1-phase state that evolved into an S-phase (DNA-replication)-like state, followed by G2 (intermediate growth), and M (cell division) before returning the stable G1-phase state, mirroring normal progression through the cell cycle. Importantly, perturbing the Boolean network by deleting or adding interactions most often did not affect either the stability of the G1 state or the frequency with which other global states evolved into the G1 state. This suggests that the robustness of cell-cycle progression is in part due to the structure of its regulatory network. Another example of using expression models to explore model dynamics is an ODE model of epithelial to mesenchymal transition (EMT) in human cells lines (Hong et al., 2015). In addition to predicting the known reversibility of the transition between epithelial and mesenchymal cell populations, the model also predicts the existence of two stable intermediate states where cells express markers of both epithelial and mesenchymal cells. By perturbing regulatory interactions in the model, Hong et al. found that the

stability of these intermediate cell types depends on feedback loops between transcription factors (Ovol2 and Zeb1) and between miRNAs and transcription factors (miR34a and Snail1, miR200 and Zeb1). These intermediate states are of particular interest as certain human cancers display characteristics of both epithelial and mesenchymal cells (Hong et al., 2015). In general, dynamic models like these offer the advantage of being able to perturb complex systems *in silico* to guide or supplement experimental approaches.

Finally, expression models have been used to discover important features of gene regulatory systems by looking at differences in performance after including/excluding different features. Although not the sole focus of their study, in building thermodynamic models of genes which regulate segmentation in *Drosophila*, Segal et al. (2008) found that including CREs that were neither enriched amongst segmentation gene promoters nor expected to bind to high affinity transcription factors were nevertheless important to accurately predict expression. These weak binding sites were found to be clustered with other *cis*-regulatory sites that bind the same transcription factors, suggesting that they might play a role in cooperative binding, which is important for predicting the sharp boundaries of expression between segments that are observed in nature. Taking a different approach, Zou et al. (2011) used support vector machine to predict genes in *A. thaliana* that are differentially expressed in response to stress based on the presence of CREs in the promoter of the gene. In addition to experimentally identified CREs, Zou et al. included computationally predicted pCREs enriched in the promoters of abiotic and biotic stress-responsive genes, respectively. Including these pCREs improved the performance of the model, suggesting that they represent bona-fide binding sites for as of yet unidentified TFs. They also identified pairs of CREs enriched amongst stress-responsive genes, the inclusion of which

further strengthened the prediction of stress response. Like the results of Segal et al., this finding indicates that cooperative binding plays an important role in stress regulation.

In the following chapters, I will present three applications of modeling gene expression I employed in my research. First, I used two models of cycling expression to identify diel expressed genes in the green alga *Chlamydomonas reinhardtii* and cluster them according to the timing of peak expression or phase. pCREs enriched in each phase-cluster were identified and subsequently used to predict the expression phase of diel genes. In the next chapter, I further explored predicting cyclic expression by comparing the performance of four different sets of regulatory interactions defined based on experimental evidence in predicting cell-cycle expression in *S. cerevisiae*. I also looked how the prediction performance was affected by including network motifs such as feed-forward loops as features and combining the best features from multiple data sets. Known cell-cycle regulators were identified as being amongst the most important TFs for correctly predicting cell-cycle expression. However, interactions amongst TFs that were neither individually important nor annotated cell-cycle regulators were also necessary to accurately predict expression. In the final chapter, I describe a different approach to understanding expression regulation, by modeling the evolution of ancestral expression and regulatory states in duplicate pairs of TFs. A system of ODEs was used to model the loss of expression and regulatory sites between these duplicate TFs, and this model suggests that asymmetry between copies, were one duplicate retains ancestral states and other diverges, is favored. Together these studies illustrate how expression modeling can be applied to a wide variety of biological questions as well as answer questions about how cyclically expressed genes are regulated and how the GRNs that control such complex patterns of expression may have evolved.

CHAPTER 2: PREVALENCE, EVOLUTION, AND *CIS*-REGULATION OF DIEL TRANSCRIPTION IN *CHLAMYDOMONAS REINHARDTII*[1]

[1] The work described in this chapter has been published in the following manuscript

Nicholas Panchy, Guangxi Wu, Linsey Newton, Chia-Hong Tsai, Jin Chen, Christoph Benning, Eva M. Farre, Shin-Han Shiu (2014) Prevalence, Evolution, and cis-Regulation of Diel Transcription in Chlamydomonas reinhardtii. *G3* **4**:2461-2471

ABSTRACT

Endogenous (circadian) and exogenous (e.g. diel) biological rhythms are a prominent feature of many living systems. In green algal species, knowledge of the extent of diel rhythmicity of genome wide gene expression, its evolution, and its cis-regulatory mechanism is limited. In this study, we identified cyclically expressed genes under diel conditions in *Chlamydomonas reinhardtii* and found that ~50% of the 17,114 annotated genes exhibited cyclic expression. These cyclic expression patterns indicate a clear succession of biological processes during the course of a day. Among 237 functional categories enriched in cyclically expressed genes, >90% were phase-specific, including photosynthesis, cell division and motility related processes. By contrasting cyclic expression between *C. reinhardtii* and *Arabidopsis thaliana* putative orthologs, we found significant but weak conservation in cyclic gene expression patterns. On the other hand, within *C. reinhardtii* cyclic expression was preferentially maintained between duplicates and the evolution of phase between paralogs is limited to relatively minor time shifts. Finally, to better understand the *cis* regulatory basis of diel expression, putative *cis*-regulatory elements were identified that could predict the expression phase of a subset of the cyclic transcriptome. Our findings demonstrate both the prevalence of cycling genes as well as the complex regulatory circuitry required to control cyclic expression in a green algal model, highlighting the need to consider diel expression in studying algal molecular networks and in future biotechnological applications.

INTRODUCTION

Diel (24 hour, day/night periods) cycles dictate physiological changes at different times of day in many organisms. The timing of these physiological oscillations is regulated by a combination of environmental, metabolic and circadian signaling processes (Farre 2012; Kinmonth-Schultz et al. 2013; Song et al. 2013; Fonken and Nelson 2014). For example, circadian clock mutants lead to phase changes under entrained diel conditions (i.e. light-dark cycles) and changes in photoperiod sensitivity (Yanovsky and Kay 2002; McNabb and Truman 2008). Oscillations can be a direct adaptation to environmental cycles, for example restricting photosynthesis and protection against UV radiation to periods of light. Diel cycles also influence biotic responses such as defense mechanisms (Arimura et al. 2008; Goodspeed et al. 2012; Baldwin and Meldau 2013) and mutualistic interactions (Frund et al. 2011; Lehmann et al. 2011). Mechanistically, many of these cycling responses are regulated at the transcriptional level. For example, in the green alga *Chlamydomonas reinhardtii*, oscillations in starch levels are partially regulated by the cyclic expression of ADP-Glucose pyrophosphorylase (Ral et al. 2006). However, some circadian regulated processes are controlled at the post-transcriptional level (Kojima et al. 2011) and/or by the interaction between transcriptional and post-translational regulation (Kinmonth-Schultz et al. 2013; Song et al. 2013). Early transcriptome analyses of three model organisms, *Arabidopsis thaliana, Drosophila melanogaster,* and *Mus musculus*, indicated that between one and ten percent of genes exhibit circadian oscillation with periods of ~24 hr (Doherty and Kay 2010). Moreover, in photosynthetic organisms, 30-90% of genes cycle under diel conditions (Michael et al. 2008; Monnier et al. 2010; Shi et al. 2010; Filichkin et al. 2011). In land plants, about a third of the genes that cycle in light/dark are also circadian regulated (Michael et al. 2008; Filichkin et al. 2011). Several *cis*-regulatory elements (CREs)

necessary for circadian regulated gene expression have been identified (Michael and McClung 2002; Harmer and Kay 2005; Michael et al. 2008), although it remains an open question how well the identified CREs explain global cyclic expression patterns.

The green alga *C. reinhardtii* has been used extensively to study physiological processes under the control of circadian and/or diel cycle (Mittag et al. 2005; Matsuo and Ishiura 2010). *C. reinhardtii*'s size, short life-cycle, and extensive genetic tool set make it an ideal model organism (Harris 2001) particularly for studies such as experimental evolution from single to multi-cellularity (Ratcliff et al. 2013) and the genetic engineering of triacylglycerol accumulation in algae (Grossman et al. 2007; Hu et al. 2008; Siaut et al. 2011). *C. reinhardtii* has also been used to study rhythmic responses to light (Bruce 1970), ammonium (Byrne et al. 1992) and nitrogen availability (Pajuelo et al. 1995). However, studies of cyclic expression in *C. reinhardtii* have been limited to single (Mittag et al. 2005; Matsuo and Ishiura 2010) or relatively small sets of genes (Kucho et al. 2005). Despite the large evolutionary distance, there are some conserved elements between both the circadian (Corellou et al. 2009; Matsuo and Ishiura 2010) and the photoperiodic (Romero and Valverde 2009) oscillators of flowering plants and green algae, raising the question whether and to what extent cyclic expression is conserved. Therefore, a genome wide analysis of cyclic expression in *C. reinhardtii* can provide insight not only into cyclic physiological behavior in green algae, but also how this behavior has evolved in divergent lineages of the *Plantae*. Such an analysis will also be relevant to economically important processes in algae such as oil production.

In this study, we examined gene expression patterns under diel conditions in *C. reinhardtii*. We characterized the prevalence of cycling gene expression in the *C. reinhardtii* genome and observed that genes involved in distinct biological processes are consistently

expressed at certain times during the day/night cycle. We also investigated the conservation of cyclic expression patterns between orthologs in *C. reinhardtii* and *Arabidopsis thaliana*, which diverged ~650-800 million years ago (Sanderson et al. 2004) and the evolution of cycling paralogous genes. Finally, to understand the cis-regulatory basis of diel expression, we identified putative CREs (pCREs) associated with cyclic expression at different phases and investigated how these pCREs can be used to predict cycling gene expression.

RESULTS AND DISCUSSION

Cycling gene expression is extensive in the *C. reinhardtii* genome

To characterize cyclic expression in *C. reinhardtii*, the expression profiles of 17,114 annotated *C. reinhardtii* genes were defined from samples taken at three hour intervals over two 24-hour time courses (see Methods). A gene was defined as cyclically expressed if it exhibited statistically significant, non-random variation at a regular period as identified by either COSPOT or DFT (see Methods). The union of predictions for both methods covered 8072 cyclically expressed genes (47.2% of the *C. reinhardtii* genome), which we hereafter refer to as "cycling genes". Both approaches generated cyclic expression models that correlated with the original expression data, with an average Pearson correlation coefficient of 0.987 for COSPOT and 0.880 for DFT. The correlation for COSPOT models is higher compared to that of DFT because COSPOT models are fit directly to the overall pattern of the data while the DFT models are based only on variations which occur at a period of 24-hours. Taken together, cyclic variation in gene expression represented the predominant form of non-linear variation in RNA content at both the genome wide and individual gene level.

Cyclic variation can be described using three parameters: period, amplitude, and phase (**Figure 2.1A**). Using the fitted models, we inferred the period, amplitude and phase of all cycling genes in the *C. reinhardtii* genome. The distribution of period for our set of cycling genes was centered around 24 hours (+/- 1.10 hours, 95% confidence interval) (**Figure 2.1B**, **Supplemental Figure 2.1A**). The amplitude of cyclic expression was highly correlated with mean expression level ($r^2 > 0.7$) and, on average, was only half the size of the mean, indicating that most cycling genes are expressed at some constitutive level even during the trough of the cycle (**Figure 2.1C, Supplemental Figure 2.1B)**. The phase distribution of cycling genes was

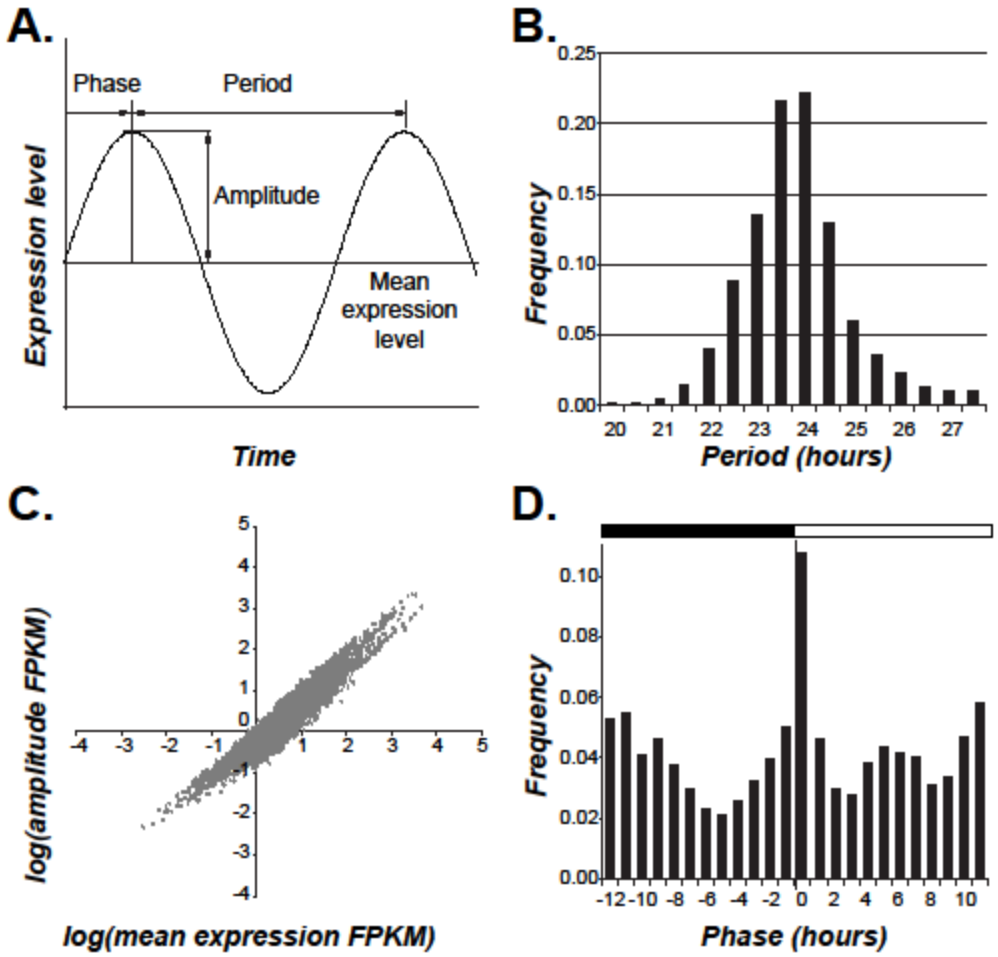


Figure 2.1 Period, amplitude, and phase of cyclic expression. (A) Three properties of cyclic variation: period, amplitude, and phase. (B) The distribution of period of cycling genes identified in *C. reinhardtii*. (C) The relationship between amplitude and mean expression level in FPKM (Fragments per Kilobase of transcript per Million mapped reads). (D) The distribution of the phase of cycling genes.

bimodal with one peak at around ZT 0 (20.6% of cycling genes) and a second around ZT 12 (16.4% of cycling genes), corresponding to the night-to-day and the day-to-night transitions, respectively (**Figure 2.1D**, **Supplemental Figure 2.1C**). Our finding concurs with the phase distribution reported for *A. thaliana* and other plant species under diel conditions (Michael et al. 2008; Filichkin et al. 2011) as well as a subset of circadian genes in C. reinhardtii (Kucho et al. 2005).

Phases of cycling gene expression are associated with a succession of biological functions

Earlier studies have shown that multiple processes in *C. reinhardtii* have specific rhythms, including the expression of key photosystem components (Hwang and Herrin 1994; Jacobshagen and Johnson 1994) and the timing of gametogenesis (Jones 1970). Thus we first asked which processes tend to be rhythmic by identifying GO terms with an over-represented number of cycling genes. We found that cycling genes were enriched in 44 GO terms, including those related to the chloroplast, photosynthesis, and ribosomal subunits (**Supplemental Table 2.1**). Among these terms, the most striking pattern was that 207 of 252 flagella related genes showed cyclic expression. In particular, 80% of cyclically expressed flagella genes (167 of 207) had peak expression at ZT 21, suggesting that biological functions can be phase specific. To further explore the association between phase and function, cycling genes were assigned to eight "phase clusters" (ZT 0, 3, 6, 9, 12, 15, 18, and 21; **Figure 2.2A**) and enrichment of GO categories within each cluster was determined.

We found that 237 GO terms had over-represented numbers of genes in ≥1 phase cluster (**Figure 2.2B).** Enrichment values for each term in each phase group can be found in **Supplemental File 2.1**. The greatest number of enriched terms was found in the ZT 21 cluster, just before the night-day transition, (40/237, 16.7%) and the ZT 9 cluster, just before the day-

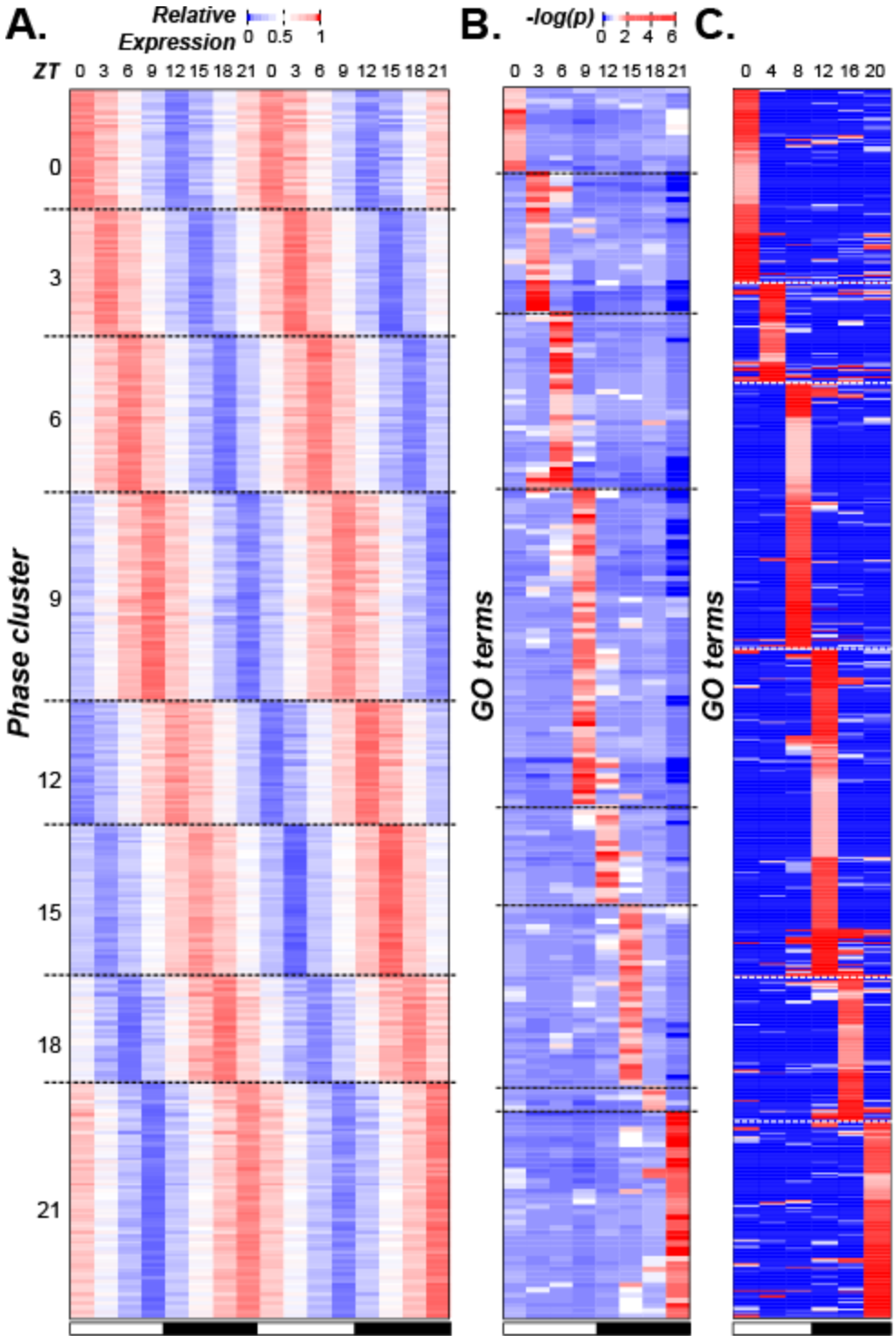


Figure 2.2 Phase of gene expression and cyclically expressed GO terms. (A) The normalized (relative) expression of each cycling gene in *C. reinhardtii* (each row) across the 48-hour period

Figure 2.2 (cont'd)

(columns). Genes were assigned to phase clusters based on the predicted time of peak expression. Genes in each phase cluster were ordered using hierarchical clustering. The white and black bars below indicate samples from the light and the dark periods, respectively. (B) The test statistics of GO term (rows) enrichment in each phase (columns) in *C. reinhardtii*. The -log(*p*-value) of the Fisher exact test is plotted. GO terms are ordered along the y-axis according to the most enriched phases and hierarchically clustered within each phase. (C) The test statistics of GO term enrichment in each phase in *A. thaliana*. Methods for assigning GO terms to phase, clustering, and the color legend are the same as in (B).

night transition (61/237, 25.7%). We also observed that over-represented GO terms tended to be phase-specific: of all 237 terms, only 19 were enriched in >1 phase and 12 of those were enriched only in two adjacent phases (**Figure 2.2B**, **Supplemental Figure 2.2A)**. In contrast, the majority of under-represented categories (51%) spanned ≥ 4 phases (**Supplemental Figure 2.2B**). Thus, genes involved in the same process not only tended to be enriched in a particular phase of expression, but were also depleted in other phases. This phase-specificity of functional categories was consistent with previous studies of light-response, metabolism, cell division, and flagellum biogenesis in *C. reinhardtii* demonstrating cyclic behavior at a specific time of the day (Jones 1970; Cavalier-Smith 1974; Teramoto et al. 2002). For example, DNA replication and mitotic events in *C. reinhardtii* are restricted to the early hours of the dark period (Jones 1970): not only is the transition into darkness required for normal cell division (Voight and Munzer 1987), but DNA replication and cell separation occur between 2-5 hours after the light-dark transition (Fang et al. 2006). This specific timing of DNA replication after the light to dark transition matches the phase of expression for cycling genes related to this process. Alternatively, the gradual increase in expression of replication associated genes towards a peak early in the dark period may track with increases in cell size, as it has been shown that the concentration of cell cycle regulatory proteins HA-MAT3, DP1, and E2F1 remain constant in spite of the increase in cell volume during G1 (Olson et al. 2010). We should note that many of the phase-specific functional categories uncovered here, such as amino acid biosynthesis, phosphorelay activity, and mRNA splicing were not previously known to show time-specific cycling behavior in *C. reinhardtii*. While correlation alone is insufficient to prove causation, the coordination between cyclic expression and function is highly suggestive that timing of transcription can regulate the timing of higher order biological processes.

Based on the apparent association between phase and function in this as well as in prior studies, GO terms were classified into broad "functional groups": (1) ribosome and translation, (2) photosynthesis and light response, (3) mitochondria and metabolism, (4) cell cycle and mitosis and (5) microtubules and flagella (**Figure 2.3**, **Supplemental Table 2.2**). We found that group 1, 2, and 3 were over-represented in the middle of the day (ZT 3 and 6), group 4 in the early and mid-night (ZT 12 and 15), and group 5 at the end of the dark-period (ZT 21) (**Figure 2.3A**). Consistent with the pattern of phase-specific enrichment of genes in different functional groups, the normalized expression profiles of cycling genes in each functional group clearly demonstrated phase specificity (**Figure 2.3B-F**). The diel expression data also highlighted the possibility of distinguishing different components of a biological process. For example, group 5 genes are involved in forming microtubules and subsequently flagella. Within this group, genes associated with the microtubule cytoskeleton peaked earlier in the dark period while those associated with flagellum assembly peaked toward the end (**Figure 2.3L**), representing a clear delineation between spindle body formation and flagellar regeneration as described previously (Cavalier-Smith 1974). Taken together, our findings suggest that the timing of biological processes (translation, cell-replication, and regeneration of the flagellum) may be determined by transcriptional regulation.

C. reinhardtii and *A. thaliana* orthologs have limited conservation in cycling gene expression patterns

To test if the functional coordination and phase specificity of cyclic expression observed in *C. reinhardtii* can be found in related multicellular species, cycling genes were identified in *A. thaliana* using the same methods and cutoff values applied to *C. reinhardtii* on an existing diel expression data (Blasing *et al.* 2005). A total of 4945 genes in *A. thaliana* were identified as

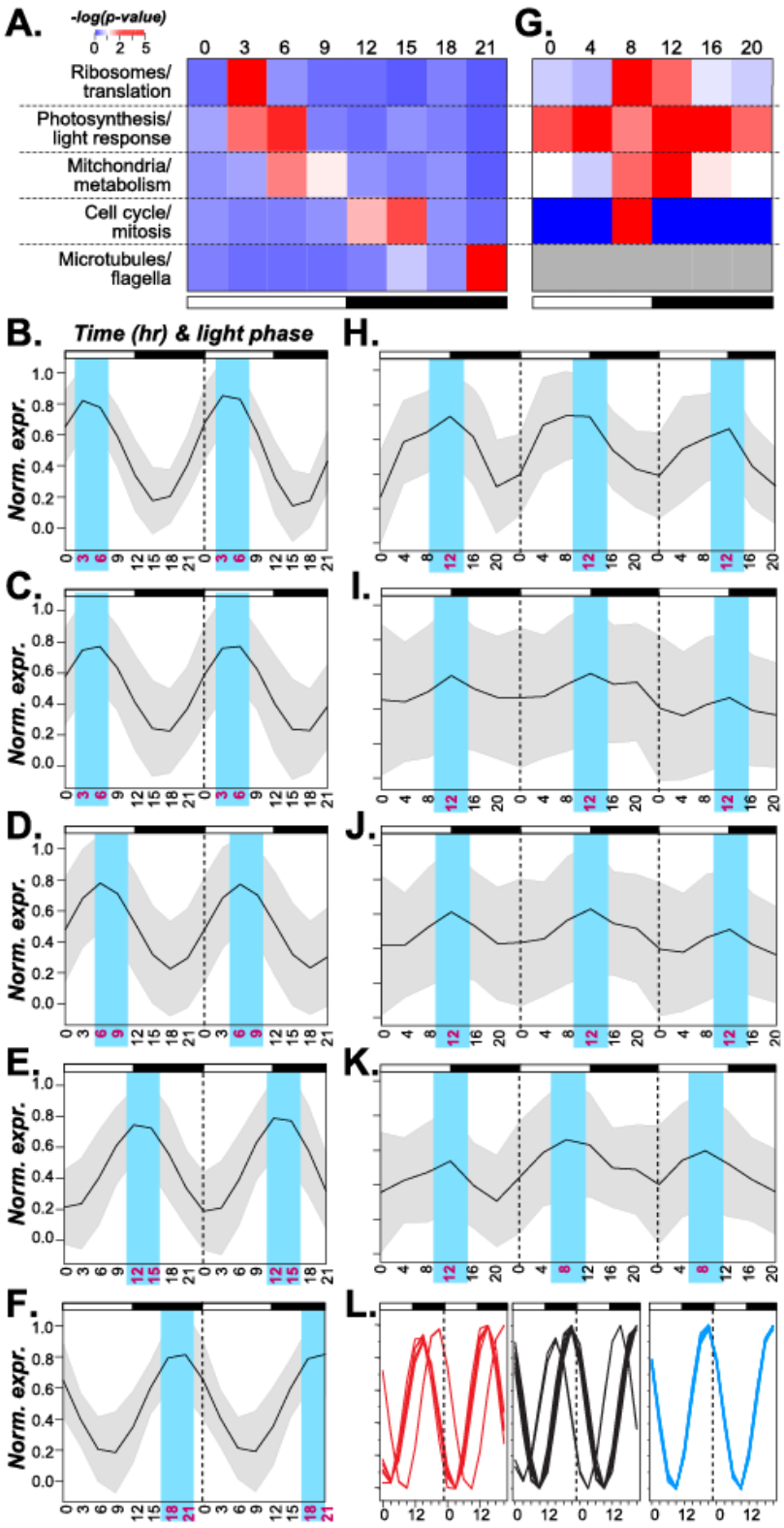


Figure 2.3 Phase specific expression of broad functional categories. (A) Enrichment test statistics in each functional group (row) and in each phase cluster (column) among *C. reinhardtii*

Figure 2.3 (cont'd)

cycling genes. The color indicates the averaged –log(*p*-value) of GO terms in a functional group (**Supplemental Table 2.2**). (B-F) Normalized expression profiles of genes in each functional group in *C. reinhardtii*. The black line indicates average expression values. The grey area represents plus/minus one standard deviation. (B) Ribosomes/Translation (C) Photosynthesis/Light-response (D) Mitochondria/Metabolism (E) Cell-cycle/Mitosis (F) Microtubules/Flagella (G) Enrichment test statistics for functional groups in *A. thaliana.* The functional group designation and color legends are the same as (A). Gray: not applicable. (H-K) Normalized expression profiles of genes in each functional group in *A. thaliana.* (H) Ribosomes/Translation (I) Photosynthesis/Light-response (J) Mitochondria/Metabolism (K) Cell-cycle/Mitosis (L) Expression profiles of genes in the microtubule cytoskeleton (red), flagellum assembly (blue) and cell projection organization (black) categories.

cycling (21.7% of the annotated genes), less than half of what was seen in *C. reinhardtii*. This difference is in part due to a lower sampling density of the *A. thaliana* data (once every 4 hours), though the overall time span covered was longer (3 days). It is also possible that the mixture of different cell types in *A. thaliana* samples could mask some rhythmic expression patterns. We also observed that 992 GO terms in *A. thaliana* were over-represented in ≥1 phases compared to 237 in *C. reinhardtii*, which is likely a function of significantly better annotation (**Figure 2.2C**). Enrichment values for each term in each phase group can be found in **Supplemental File 2.2**.

In contrast to the strict phase-specificity in *C. reinhardtii*, *A. thaliana* group 2 GO terms (photosynthesis and light response) were enriched amongst cycling genes in all six time points, but were predominant at ZT 4. The other three groups (group 1, 3, and 4) were restricted to at most two adjacent phases (**Figure 2.3G)**. Compared to *C. reinhardtii*, there is a greater variance in the phase of expression amongst the *A. thaliana* cycling genes within each group, potentially due to the fact that the *A. thaliana* expression data was derived from samples of mixed tissues and cell types. Nonetheless, the peak expression of photosynthetic, mitochondrial, and ribosomal genes occurred at a similar time, as was observed in *C. reinhardtii* (**Figure 2.3H-K**). These results suggest that cyclic expression is conserved between a subset of functionally related genes, in both unicellular and multi-cellular plant systems.

Due to the concern that the phase-specificity differences between *C. reinhardtii* and *A. thaliana* might be due to annotation quality difference, we next examined the degree to which cycling gene expression was conserved between orthologous genes in these two species. Among 11,845 putative orthologs, 1,464 (12.4%) showed cyclic expression in both species (referred to as "co-cycling" orthologs), which is significantly higher than the random expectation (Chi-Squared Test, $pv < 0.001$). The conserved co-cycling genes encode components of the ribosome

(particularly the small subunit), plasma and thylakoid membrane components, or are involved in stress response (Fisher Exact Tests, $pv < 0.05$). Nonetheless, we should emphasize that the difference between the observed and expected proportion of co-cycling orthologs was only 2.4%. Thus most cycling genes in *C. reinhardtii* are not cyclic in *A. thaliana* and vice versa. In addition, while the amplitude of cyclic expression is significantly correlated among co-cycling orthologs ($r^2 = 0.30$, $pv < 10^{-100}$), there are only weak relationships between their phases ($r^2 < 0.01$, $pv < 0.006$). The *A. thaliana* and *C. reinhardtii* lineages diverged 650-800 million years ago (Sanderson *et al.* 2004) and have extensive differences in life histories, distribution, complexity, and physiology. Thus the conserved components of cyclic expression are likely core processes strongly selected to be maintained, including photosynthetic, mitochondrial, and ribosomal genes (**Figure 2.3H-K**). However, most orthologs between green algae and flowering plants have divergent patterns of cyclic expression, and the extent of cyclic expression divergence highlights the fact that cycling gene expression can be plastic.

Conservation of cyclic expression is more prevalent amongst older duplicate genes

To further assess how quickly cyclic expression divergence occurred, we asked how the pattern of cycling gene expression evolved between duplicated genes in *C. reinhardtii*. Gene trees were inferred based on similarity of known protein domains and we retained only the closet pairs of paralogs (i.e. those separated by only a single ancestral node) for subsequent study (see Methods). The frequency with which the pattern of gene expression (cycling or non-cycling) was identical or divergent was compared to the timing of the inferred duplication event, estimated using the synonymous substitution rate (*Ks*) (**Figure 2.4A**). The overall frequency of diverged duplicates (one paralog cycling, the other non-cycling) increased with *Ks*, approaching an asymptote of ~0.45 for $Ks > 0.9$. While the frequency of non-cycling duplicates decreased with

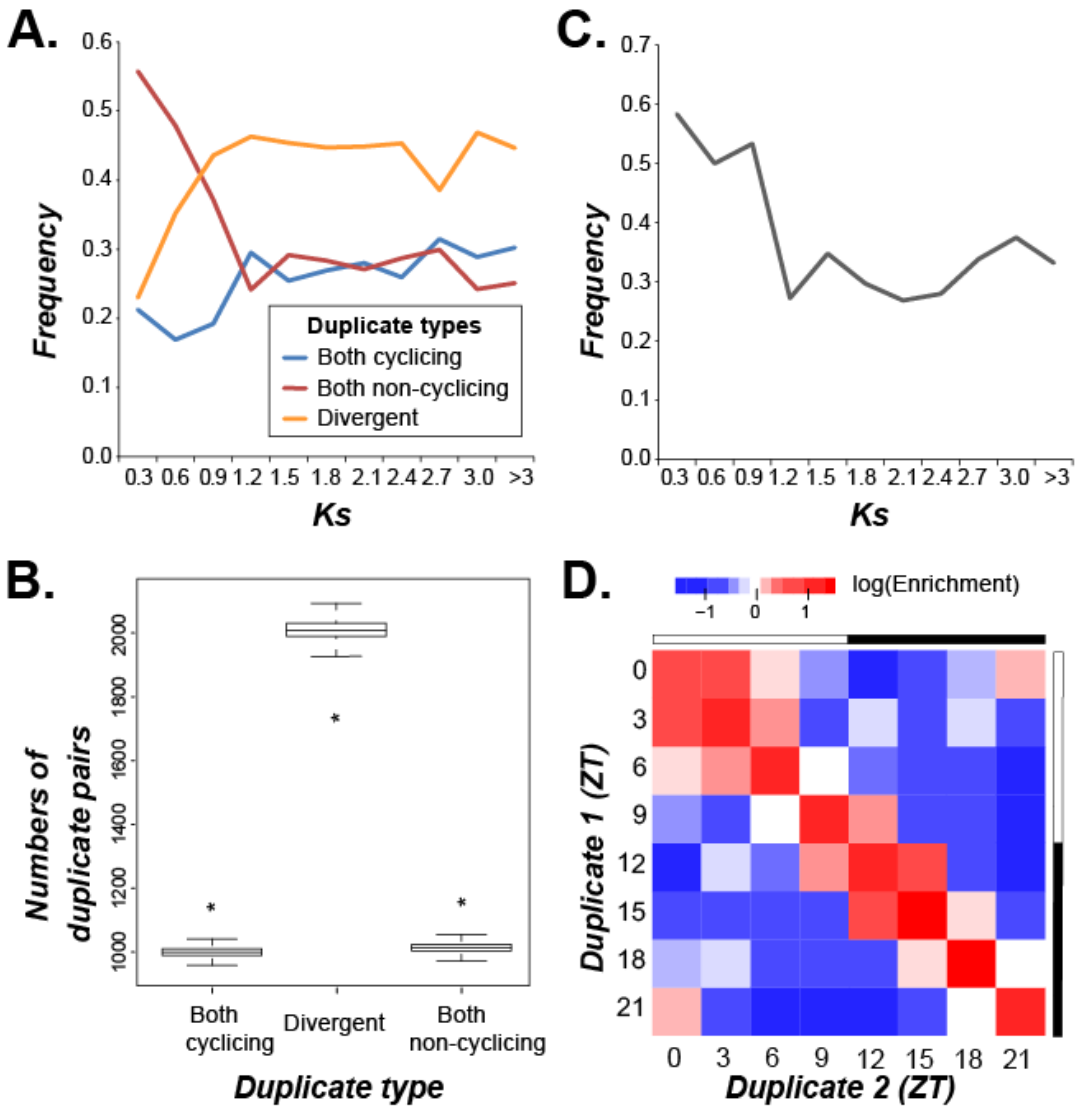


Figure 2.4 Conservation of cyclic expression and phase of cyclic expression. (A) The frequency at which duplicate pairs of genes in *C. reinhardtii* maintain cycling expression, maintain non-cycling expression, or diverge as a function of the synonymous substitution rate (Ks). (B) Distribution of cycling retention, non-cycling retention and divergence between duplicate pairs in random simulations. The black bars cover the inter-quartile range of each distribution, and error bars represent the 95% confidence interval. Observed values are indicated by asterisks. (C) The frequency at which the phase is retained in pairs of cycling duplicates as a

Figure 2.4 (cont'd)

function of Ks. (D) Enrichment values for phase retention (diagonal values) and phase change (off diagonal values) between actual duplicates and duplicate pairs in random simulations.

Ks, the frequency of cycling duplicates was greater on average for *Ks* > 0.9 indicating a net gain of cycling expression as duplicates age. We hypothesized that this gain of cycling expression results from a bias in the rate at which duplicate genes diverge that favors the cycling state.

To test this hypothesis, we examined if the observed changes in the frequency of retention can be explained without assuming different rates of divergence. Therefore, a null model of duplicate gene divergence was created using a system of difference equations (see Methods). We fit the transition probabilities using the difference in frequencies between *Ks* 0.6 and 0.9, and the predicted frequencies of identical and divergent duplicates closely matched our observed results at all time points (root mean squared error = 0.03), showing the same pattern of increases and decreases (**Supplemental Figure 2.3**). Hence, we have no evidence of a differential rate in divergence between cycling and non-cycling duplicates, however the predicted probability of transition from identical to divergent (0.42) is less than the probability of transition from divergent to identical (0.53), suggesting that there is a preference for maintaining duplicates in an identical state. This is consistent with our finding that the observed frequency of paralogs with identical states tends to be significantly higher than expected under random association (Z-test, pv < 10^{-17}; **Figure 2.4B**). In contrast, the frequency of paralogs with divergent state is significantly lower than expected (**Figure 2.4B**).

Next, we examined the frequency with which phase is identical amongst pairs of the cycling duplicates. Overall, the number of co-cycling paralogs for which the phase of cyclic expression was identical is more than twice the number randomly expected (Z-test, *pv* < 10^{-85}) with 33.7% of co-cycling duplicates sharing the same phase. The identical phase state was more common amongst cycling duplicates with lower *Ks* and there was a sharp decrease in the frequency of duplicates with identical phases going from a *Ks* of 0.9 to 1.2 (**Figure 2.4C**). Next

we explored if there was a bias in the magnitude of phase change between co-cycling duplicates (**Figure 2.4D**). We found that small phase divergences of +/- 3 hr (covering 28.3% of all duplicates) tended to be enriched relative to random expectation, in particular at ZT0/ZT3 and ZT12/ZT15, although the identical phase state is still the most highly enriched scenario. Additionally, there was an inverse, linear relationship between the magnitude of the difference in phase between cycling duplicates and the enrichment of phase-shift events relative to random expectation (all cycling duplicates, $r^2 = 0.91$; duplicates with $Ks > 0.9$, $r^2 = 0.93$), indicating that large differences in phase between duplicates occur less frequently than expected by random chance. Furthermore, we found 33 GO terms enriched (adjusted p-value < 0.05) amongst cycling duplicates with the same phase, the majority of which (88%) were previously found to be enriched in a specific phase of cyclic expression.

Cycling genes are enriched for specific putative Cis-regulatory elements

The coordinated expression of functionally related genes suggests the existence of one or more regulatory mechanisms that drive phase specific expression. While mRNA levels may be affected at multiple levels of regulation, we chose to focus on transcriptional regulation driven by *cis*-regulatory sequences as circadian rhythm related *cis*-elements have previously been identified in plant and animal models (Michael and McClung 2002; Ueda et al. 2005; Michael et al. 2008). Using a motif finding pipeline (Zou et al. 2011), we found 687 putative *cis*-regulatory elements (pCREs) in the 1kb regions upstream of the transcriptional start sites of cycling genes for each of the eight *C. reinhardtii* phase clusters (Fisher Exact Test, adjusted $pv < 0.05$). The top enriched motifs for each phase can be found in **Figure 2.5**, and the entire list of enriched motifs can be found in **Supplemental File 2.3**. Each phase had 60-84 associated pCREs, except for ZT 15 with 169; however, more than 20% of pCREs (141/687) were enriched in >1 phase and 43.8%

of ZT 15 pCREs (74/169) were enriched in ≥1 other phases (mostly ZT 12; **Figure 2.6A**). Therefore, each pCRE was assigned to the phase cluster in which it was most significantly enriched.

To further assess whether the pCREs are meaningful, they were used to establish classifiers to predict cyclic expression in different phases. First, the pCREs assigned to each phase were used to predict which genes are cyclic in a naïve manner. That is, for pCREs enriched in a particular phase, we simply predicted that all genes with ≥1 pCREs mapped to their promoters would cycle at that phase. The performance of these predictions was assessed using the area under the receiver operating curve (AUC-ROC), a metric which quantifies the ability of a method to predict positive examples which, in our case, is phase specific expression. Perfect predictors have an AUC-ROC of 1 while random guessing has a value of 0.5; our naïve classification of phase had AUC-ROCs that ranged from 0.58 (ZT 9) to 0.62 (ZT 12) indicating that this simple classification procedure performed marginally better than randomly assigning phase (**Figure 2.6B**). The same conclusion can be reached based on the F-measure, another prediction performance metric (**Figure 2.6C**). Next, to further improve the prediction of the phase of cyclic expression, we used the support vector machine (SVM) algorithm to classify cycling genes according to the presence or absence of all pCREs (see Methods). The SVM classifier shows improved performance compared to naïve classification (**Figure 2.6B-C, Supplemental Table 2.3**) but AUC-ROC values are still relatively low, ranging from 0.58 (ZT 9) to 0.65 (ZT18) (**Supplemental Figure 2.4)**. We also identified two pCRE association rules enriched in specific phases of cyclic expression using CBA (Liu et al. 1998); however, adding these rules to the SVM prediction models did not significantly improve the overall predictive power of our pCREs as the AUC-ROC increased by at most 0.01.

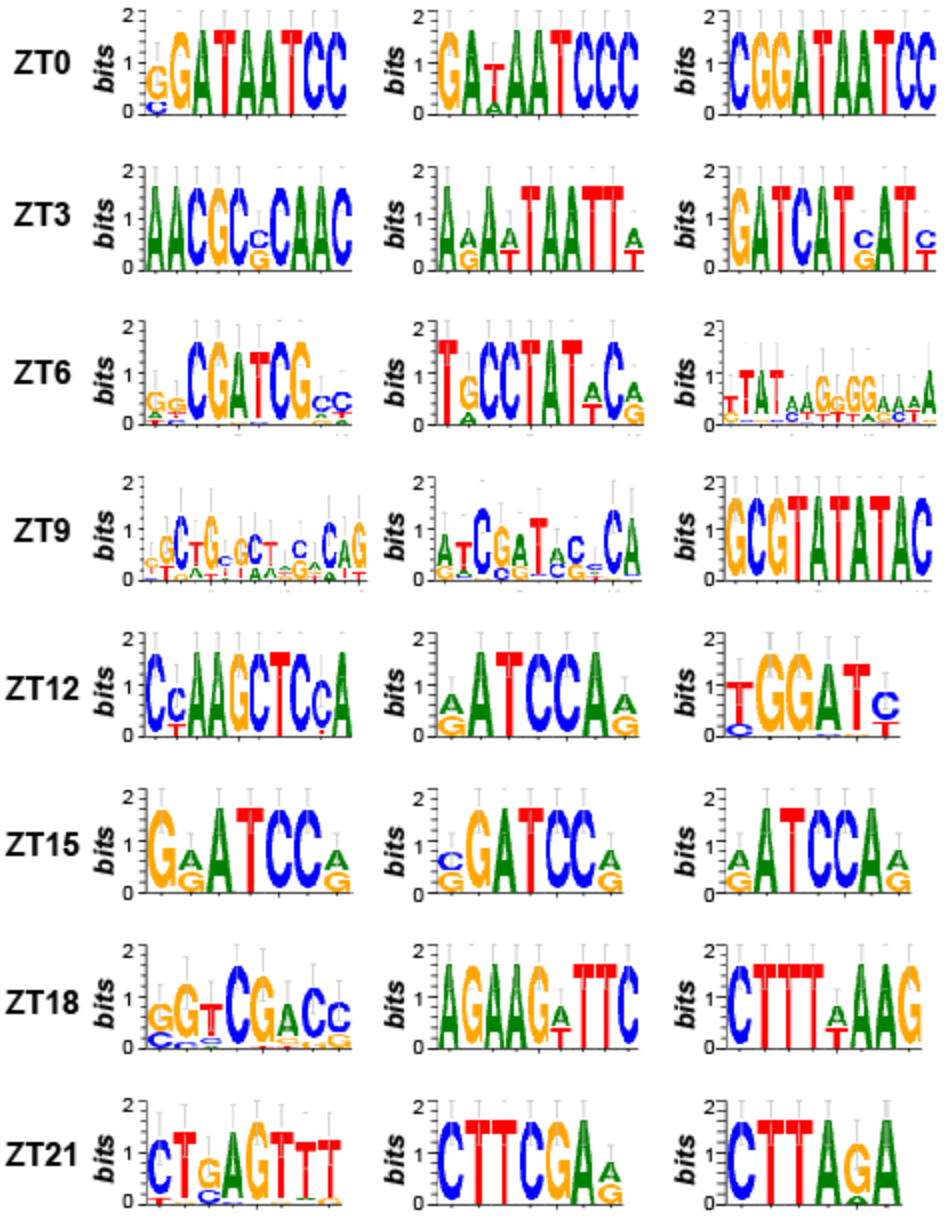


Figure 2.5 Top three pCREs enriched in each phase cluster of cyclic genes. Sequence logos representing the top three putative *cis* regulatory elements (pCREs) enriched in each phase cluster of cycling genes in *C. reinhardtii*.

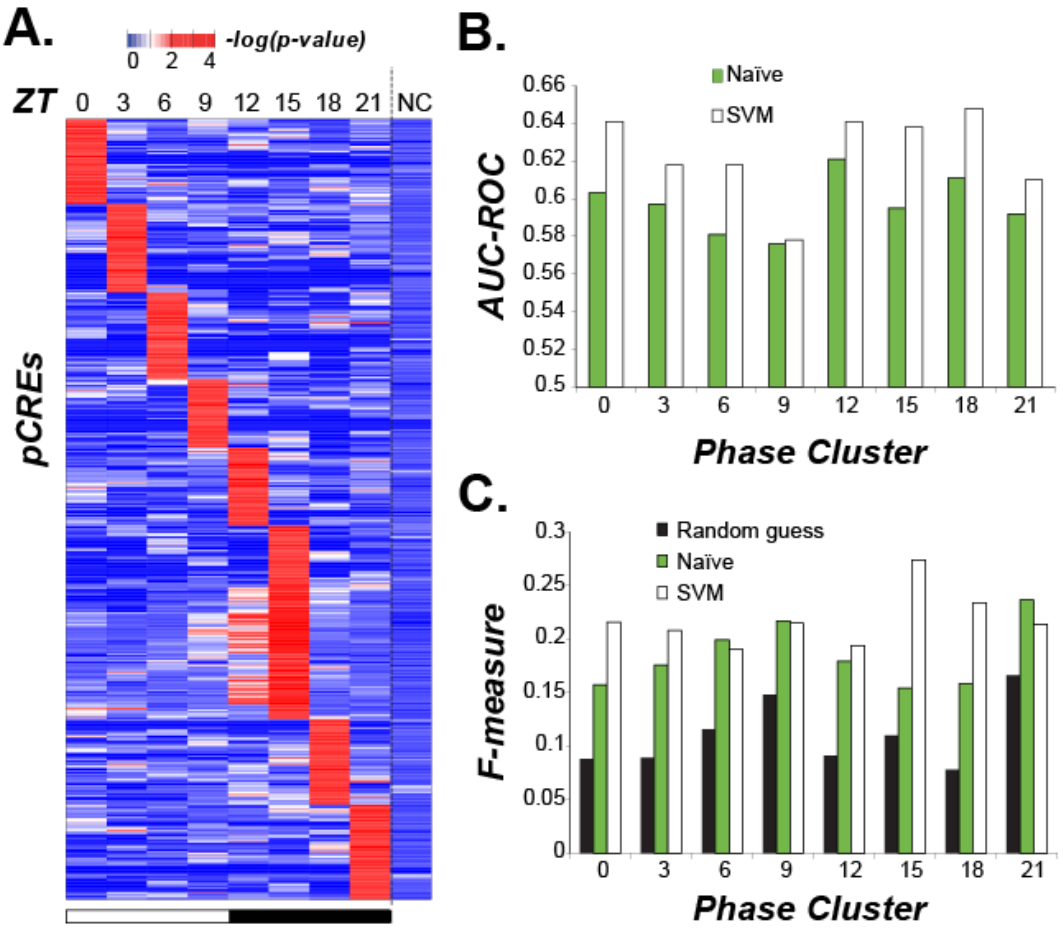


Figure 2.6 Enrichment and performance of phase-specific pCREs. (A) The enrichment test statistics of 687 pCREs (rows) in genes of each phase cluster and non-cyclic (NC) genes (columns). (B) The area under the curve of the receiver operating characteristic (AUC-ROC) for phase expression prediction with naïve (green) and Support Vector Machine (SVM, white) classifiers. (C) The F-measures for phase expression prediction based on random guess (black), naïve (green) and SVM (white) classifiers.

Given that the *C. reinhardtii* pCREs are computationally derived, we next asked how well a known, experimentally verified, phase-specific *cis*-regulatory element may predict cyclic expression. For this purpose we examined the Evening Element that is necessary and sufficient to drive circadian expression in *A. thaliana* (Michael and McClung 2002; Harmer and Kay 2005; Michael et al. 2008). Using motifs related to the Evening Element identified in Michael et al. (2008), we generated a cycling gene classifier to predict the phase of the 4,945 *A. thaliana* cycling genes introduced in the earlier phase-specificity comparison section. The optimal AUC-ROC of the Evening Element classifier was 0.57 and 0.56 at ZT 0 and 12 hours, respectively (compared to 0.58-0.65 in *C. reinhardtii* pCRE predictions). Therefore, although the Evening Element is known to function as a circadian regulator, similar to *C. reinhardtii* pCRES, it has only limited predictive power on a genome wide scale. To obtain accurate predictions the presence or absence of pCREs needs to be supplemented with additional information regarding the regulation of cycling expression.

Phase of cyclic expression can be predicted for groups of genes with common expression patterns or common function

The weak predictive power of pCREs likely results from an underlying complexity in the regulation of the phase of cyclic expression, either in the form of additional control mechanisms or the existence of more discrete regulatory groups. Timing of cyclic expression may be modified by interactions amongst regulatory motifs or post-transcriptional mechanisms. It is also possible that our phase clusters might consist of multiple regulatory subgroups. To address the latter possibility, we further classified genes in each phase group into sub-clusters containing genes with highly similar expression profiles (phase-expression clusters). Using SVM, 28 of 190 phase-expression clusters covering 584 genes (7.23% of cycling genes) could be classified with

an AUC-ROC > 0.7 (these clusters are described in **Supplemental File 2.4**), which is better than any individual phase alone. The best predicted phase-expression clusters do not necessarily have stronger cyclic signals (**Figure 2.7A**) compared to the worst predicted (**Figure 2.7B**). Additionally, we eliminated size ($r^2 = 0.15$) and the correlation of expression profiles within each phase-expression cluster ($r^2 < 0.01$) as possible variables explaining the observed variance in AUC-ROC (**Supplemental Figure 2.5**). These results suggest that phase specific regulation does occur at the *cis*-regulatory level for particular groups of cycling genes and that presence or absence of pCREs alone is sufficient to accurately predict the pattern of phase specific expression for these clusters. Those pCREs which were informative (i.e. had the non-zero weights) when predicting the 28 best phase-expression clusters are listed in **Supplemental File 2.5**.

In addition to using highly similar expression patterns as a way of subdividing phase clusters, we looked for evidence of phase specific regulation amongst groups of genes in the same phase cluster that had related annotated function (phase-function clusters). Among 71 phase-function clusters, genes belonging to 12 of these clusters could be classified with an AUC-ROC > 0.7. These clusters covered 12.2% (175/1434) of genes present in all phase-function clusters, a higher percentage than the phase-expression clusters, although they constitute a smaller portion of all cycling genes due to limited GO annotation in *C. reinhardtii* (these cluster are described in **Supplemental File 2.6**). Genes in most of these functional groups displayed a clear cyclic signal (**Figure 2.7C**), except for the groups related to the nucleolus and cell wall, which were predominantly non-cyclic genes but had a statistically significant subset of phase-specific genes. Amongst the best classified sub-clusters contained genes relating to the large cytosolic ribosomal subunits (AUC-ROC = 0.73), cilium (0.72), small cytosolic ribosomal

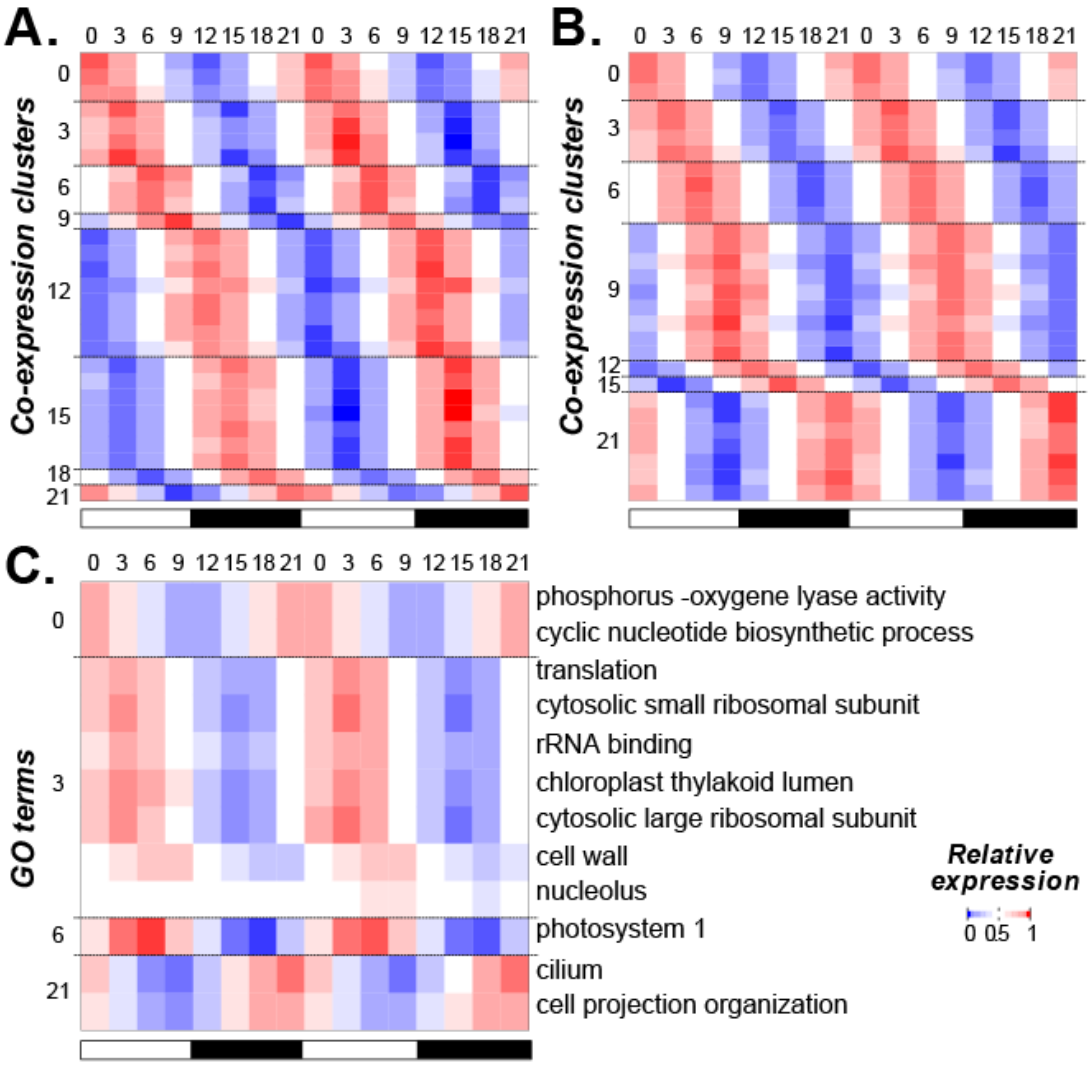


Figure 2.7 Expression of best predicted co-expression cluster and GO terms. (A) Averaged, normalized expression profile of genes in the top 28 co-expression clusters whose phase of expression can be predicted with AUC-ROC > 0.7. (B) Averaged, normalized expression profiles of genes in the bottom 28 co-expression clusters whose phase of expression can be predicted with AUC-ROC < 0.56. (C) Averaged, normalized expression profiles of genes in the 12 GO terms whose phase of expression can be predicted with AUC-ROC > 0.7. Both cycling and non-cycling genes annotated in each GO term are included.

subunit (0.72), translation (0.71), and the chloroplast (0.68). This supports our earlier observation that the cyclic patterning of large scale processes such as photosynthesis, translation, and motility may be regulated at the transcriptional level. The pCREs which had non-zero weights when predicting the 12 best phase-function clusters are listed in **Supplemental File 2.7**.

CONCLUSIONS

We have determined that cyclic expression is prevalent in the *C. reinhardtii* genome, and nearly half of all annotated genes cycle under diel conditions. There is a strong link between rhythmic patterns at the molecular and physiological levels. Diel cycling expression is influenced both by environmental factors, such as the availability of light, and endogenous factors, including metabolism and the circadian clock (Farre 2012; Kinmoth-Schultz et al. 2013; Song et al. 2013; Fonken and Nelson 2014). While the importance of photoperiod can be inferred for light-dependent (i.e. photo-synthesis) and light-sensitive (i.e. DNA replication) processes, for most cycling related functions it remains an open question as to what extent each factor influences cycling expression. This is particularly true of functions which were not previously known to exhibit cycling expression in green algae, for example, the regulation of RNA processing and amino-acid synthesis.

In addition to the relationship between cyclic expression and gene function, we found that cyclic expression was significantly conserved between paralogous genes. The proportion of divergent duplicates reaches an asymptote at $Ks > 0.9$, which is similar to what was previously observed for stress responsive duplicate genes (Zou et al. 2009). However, while there appears to be a clear preference for the partitioning of ancestral expression states in stress responsive genes (Zou et al. 2009; Dong and Adams 2011), we found that duplicates genes tend to share the same expression state with respect to cycling and that cycling duplicates preferentially retain the same or similar phase of expression. We hypothesize this pattern of cyclicity/phase conservation among duplicates points to a fundamentally distinct regulatory logic from that of stress response. In stress response, a duplicate which has lost response to one condition may still be responsive to other conditions and thus retained. However, either loss or gain of cyclicity in a duplicate gene

would mean it is no longer temporally in sync with other genes in the processes which it was originally involved in. For example, if a replication initiation factor duplicate was not in sync with the expression of other components of the replication machinery, the duplicated factor would not be functional and eventually eliminated from the genome. This argument may also apply to the conservation of phase among duplicate cycling genes. Indeed, we found that most GO terms enriched amongst co-cycling duplicates with the same phase were highly phase specific, including those associated with DNA replication and flagellar components.

Based on prior studies of stress response genes (Zou et al. 2009), we expected that the conservation of cycling expression state, particularly the phase of expression, would be correlated with the presence of shared *cis*-regulatory elements. However, contrary to this expectation, the set of putative *cis*-regulatory elements enriched in cycling genes does not accurately distinguish phase expression. While our results suggest that *cis*-regulation plays a significant role in controlling cyclic expression in *C. reinhardtii,* the presence or absence of promoter elements alone was insufficient to fully explain the observed patterns of cyclic variation across the entire *C. reinhardtii* genome. This suggests that additional regulatory components are involved in controlling cyclic expression. In other organisms the combinatorial interactions amongst regulatory factors play an important role in controlling the phase of cyclic gene expression (Harmer and Kay 2005; Ueda et al. 2005), but in *C. reinhardtii* there is evidence that response to changing light levels is mediated by multiple copies of the same or similar promoter elements (von Gromoff et al. 2006). While we did not see significant improvement when rules considering combinatorial relationships between pCREs were included in our model, this may be due to the fact that we were able to explore only a subset of all possible combinatorial interactions in our pCRE set. Additionally, post-transcriptional regulation has been

implicated in regulating circadian processes in *Neurospora crassa, A. thaliana*, and *D. melanogaster* (Kojima et al. 2011). In *C. reinhardtii*, the over or under expression of the RNA-binding protein CHLAMY1 is known to result in the disruption or loss of circadian rhythms (Iliev et al. 2006). Further studies incorporating post-transcriptional regulatory features will be necessary to improve the prediction of phase specific cyclic expression

The inability of pCREs to classify phase specific cycling expression on a genome wide scale does not contradict prior observations that certain *cis*-elements are necessary for cycling expression (Michael and McClung 2002; Harmer and Kay 2005; Michael et al. 2008). Rather it suggests that *cis*-elements alone are insufficient to explain the variation in cycling expression on a genome wide scale and that additional regulatory components remain to be discovered. Post-transcriptional regulatory mechanisms and chromatin state are two promising avenues of investigation which, in conjunction with the *cis* elements we have identified, could be used to better predict the state of cycling expression. Although there remains substantial room for further improvement, our findings contribute to a better understanding of both the function and evolutionary origins of cyclic expression in a green alga, laying the foundation for further molecular dissection of the relationships between the rhythmic gene expression and physiological functions for potential biotechnological applications.

MATERIAL AND METHODS

Growth of Chlamydomonas reinhardtii Cultures

C. reinhardtii dw15.1 was grown in TAP (Tris-Acetate-Phosphate) media in flasks without aeration, shaken at 100 rpm, at 22 °C. While the acetate present in this media provides an alternative source of carbon, allowing for *C. reinhardtii* to grow in the dark, prior studies have shown that that the cell cycle (Voight and Munzner 1987; Davies and Grossman 1994) and other metabolic cycles (Ral et al. 2006) are still synchronized in *C. reinhardtii* grown in acetate-containing media under light/dark cycles. Additionally, the amplitude and phase of cell cycle gene expression in our study and in previous studies where cultures were grown under autotrophic conditions (Bisova et al. 2005) are similar (**Supplemental Figure 2.6**). An initial 200 mL culture was grown to a density of 25 million cells mL^{-1} in constant light (50 μmol s^{-1} m^{-2}) and used to set up 50 mL cultures of 0.5 million cells mL^{-1} that were transferred to 12 hr light (50 μmol s^{-1} m^{-2}) and 12 hr dark conditions for 48 hours prior to sampling. Two biological replicates were collected every 3 hours between ZT (Zeitgeber Time, hours since last dawn) 0 and ZT 21. Each sample originated from an independent 50 mL culture. Samples collected during the light to dark or dark to light transition were taken just prior to change of conditions. For collection, 2 mL of the culture was placed in a 2 mL tube and centrifuged at max speed in at 4°C for 10 min. Amber tubes were used for samples collected during the dark period and the supernatant was removed under weak green light. The pellets were snap frozen in liquid nitrogen. The frozen samples were ground using the Qiagen tissue lyser for RNA extraction.

RNA-sequencing

RNA was extracted using the Omega eZNA Plant RNA kit. The RNA was eluted in 50 μL DEPC-H_2O and the concentration was measured using a Nanodrop (Thermo-Fisher). A portion of the RNA was diluted to 1 ng uL^{-1} to check the RNA Integrity Number (RIN) with a Bioanalyzer (Agilent). All samples had a RIN equal to or greater than 7. Library preparation and sequencing was performed at the MSU-Research Technology Support Facility using the Illumina Tru-Seq Stranded kit with an Illumina HiSeq 2500. Eight samples were sequenced in each lane using a custom bar-coding, but the two biological replicates from the same time point were run in separate lanes. The average number of RNA-Seq reads per sample was 1.81e7 and they ranged between 7.07e6 and 2.58e7. The reads from each of 16 samples (8 time points, 2 samples each time point) were mapped to the *C. reinhardtii* genome (version 4.3 from Phytozome) using Tophat (Trapnell et al. 2009) with default parameters except for intron length (min 13, max 8712) and max-multi-hits (1). Gene models on non-chromosomal fragments were not considered. FPKM (Fragments Per Kilobase of transcript per Million mapped reads) per gene was calculated using Cufflinks (Trapnell et al. *2010*) with parameter –I 8712. A high percentage of reads mapped to the genome: the least mapped sample had 82% of reads mapped and the average of all samples was 85%. Upper quartile normalization was applied to all samples to correct for technical variation as recommend in (Bullard et al. 2010). The two biological replicates were appended and used as two consecutive days for downstream analysis. Raw read data is available through the NCBI SRA, BioProject accession [PRJNA264777].

Identification of Cycling Genes

Two programs were used to identify cyclic patterns of expression in FPKM data: COSPOT (which is described in (Panda et al. 2002)) and an application of the discrete Fourier

transform (DFT). The DFT has previously been applied to the analysis of cyclic expression using RNA-Seq data (Rodriguez et al. 2013), but our method is based primarily on PRIISM (*Rosa et al.* 2012). We chose to use both COSPOT and the DFT in conjunction because we found that the combination of methods had superior coverage of known cycling genes without a substantial increase in the expected false positive rate (see **Supplemental Materials and Methods**)

In our application, we take the discrete Fourier transform of each gene expression vector in the *C. reinhardtii* FPKM data set, converting a set of 'N' FPKM values (x) in terms expression vs. time to new values (y) in terms of expression vs. frequency such that:

$$y_k = \sum_{n=0}^{N-1} x_n * e^{-i2\pi\frac{kn}{N}} \qquad (1)$$

Where x_n is the FKPM value at the nth time point and y_k is the kth frequency component with period T/k where T is the time period spanned by the expression vector. The set of frequency components represents the power spectra of the associated expression vector, that is, the contribution of each periodic cycle to the overall data. In calculating the power spectra of the expression data, we employed a non-windowed application of Welch's Method (too few data points were present tolerate the loss of information involved in windowing) to average the power spectra over subsets of the expression vector with T = 24 hr. This was done to reduce bias in the calculation of the power spectra that might be induced by a particular subset of the expression data at the cost of reducing the overall resolution of the power spectra (though this loss of resolution was primarily at the extreme ends of the spectra and should not affect our results). Furthermore, the coefficients of each power spectra were normalized prior to averaging using the following equation:

$$y_k^* = \frac{y_k - y_{\min}}{y_{\max} - y_{\min}} \qquad (2)$$

Where y_{min} is the smallest coefficient of the power spectra and y_{max} is the largest. As such, the normalized values, y_k^*, are on the interval [0,1], further reducing the affect that a single subset can have on the average power spectra. The "cyclic score" of each gene is defined as the normalized value of the 24 hour frequency component. This score is equated to a *p*-value by randomizing the order of values in each expression vector and scoring the vectors in this random population. For this study, we tested cyclic score thresholds equal to the 5th, 2nd, and 1st percentiles of the score distribution of the randomized data (equal to 0.745, 0.808, and 0.841 respectively) and chose the 2nd percentile as our cutoff for calling cycling genes (equivalent to a p-value of 0.02). In comparison, the 5th percentile of cyclic score for the set of predicted cycling genes in *C. reinhardtii* was 1. Additional information about how these thresholds were determined as well as a comparison to COPSPOT can be found in the **Supplemental Materials and Methods**.

Clustering Cycling Genes According to Phase

Cycling genes in *C. reinhardtii* were first divided by their phase of expression, that is, the Zeitgeber Time (ZT) at which peak expression occurred in the FPKM data set. Within each phase cluster, genes were ordered using hierarchical clustering implemented in R for display purposes. Phase clusters were further broken down using two-rounds of k-means clustering, implemented using custom Python scripts. K-means clustering involves initially selecting "k" random centers in parameter space and assigning genes to clusters based on their distance to the nearest center. The mean of each cluster is then used to define new centers which in turn are used to redefine clusters; this process is repeated until the clusters converge or the amount change per iteration falls below a specified threshold. The final clusters used for pCRE identification contained 10 to 90 genes. Enrichment of Gene Ontology (GO) terms and pCREs in phase groups

was done using the Fisher Exact Test and the resulting *p*-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg method (Bengamini and Hochberg 1995).

Conservation of cyclic expression and phase of expression amongst duplicate genes

Gene trees in *C. reinhardtii* were defined using the pipeline described in Zou et al. (2009) using a set of protein domains defined using PFAM (Punta et al. 2012). These domains were extracted from protein sequences and aligned using MAFFT (Katoh et al. 2002), and a phylogeny was inferred using RAxML (Stamatakis 2006) with parameters -f d -m PROTGAMMAJTT. Large domain families were divided by building neighbor joining trees with PHYLIP (Felesenstein 2005) and cutting at a distance to root ≥ 0.05 to create sub-clusters between 4 and 300 genes in size. Domains were mapped back to *C. reinhardtii* genes to infer gene trees. The gene trees, including the divided trees for large domain families, were reconciled with an existing species tree (Moreau et al. 2012) using NOTUNG (Chen et al. 2000). An archive of these gene trees in Nexus (.nex) format has been included as **Supplemental File 2.8**. Branches containing *A. thaliana* and *C. reinhardtii* genes were extracted from the overall tree. The significance of the retention rate of cyclic expression and the phase of cyclic expression was determined by randomly pairing genes in the set of duplicates 100,000 times and comparing retention among actual duplicates to the random population.

Modeling cycling state divergence of duplicate genes

The divergence of duplicate genes was modeled using a system of three difference equations with a common rate ‘d’ for the divergence of both cycling and non-cycling duplicates and a common rate ‘s’ for the reversion of diverged duplicates back to an identical state. Duplicate gene pairs were binned according to *Ks* (width = 0.3), and we assume that the initial frequency of duplicates was the same within each bin (If the initial conditions were significantly

different, we would expect to see deviation from the observed frequencies in the model predictions, which was not the case). We then solved for values of 'd' and 's' using the observed change between consecutive bins, arriving at a solution with the same qualitative behavior as the observed data. A detailed description of the model can be found in the Supporting Information.

Identification of putative cis-regulatory elements and phase prediction

Identification of pCREs in the promoter regions of *C. reinhardtii* genes followed the pipeline described in Zou et al. (2011). Cycling genes were clustered according to phase and expression profile as previously described. For each cycling gene the promoter region, defined as the first 1kb upstream of the transcription start site less any bases which overlap with another gene, was isolated. Six motif finders, AlignAce, MDscan MEME, Motif Sampler, Weeder, and YMF, were used to identify motifs enriched in the promoter region of each phase cluster compared to the promoters of all cycling genes. The resulting motifs were merged using UPGMA to reduce the number of motifs and remove redundant motifs. Merged motifs were mapped back to the *C. reinhardtii* genome using a threshold *p*-value of 1e-05.

The presence or absence of pCREs was used to predict the phase of expression of cycling genes using a Support Vector Machine (SVM) implemented in Weka (Hall et al. 2009). Given a test-set of positive and negative examples defined using n-variables (in this case, presence or absence of pCREs), SVM seeks to define a linear classifier (i.e. a hyperplane in variable space), which best divides positive and negative examples. This classifier is then used to assign subsequent data points to either the positive or negative set. A grid search of two parameters, the minimum distance between positive and negative groups (C) and the ratio of negative to positive examples in the training set (R) were used to optimize separation and pick the best classifier. The tested range of each parameter was as follows: C = (0.01, 0.1, 0.5, 1, 1.5, 2.0) and R = (0.25, 0.5,

1, 1.5, 2, 2.5, 3, 3.5, 4). Results were validated using 10-fold cross validation, which involved dividing positive and negative examples for each phase into training test sets using stratified random sampling. Each of the 10 test sets was classified by an independent SVM run and the average of the 10 runs was used to score the performance of the parameter set.

Identifying groups of genes with common expression or common function

Cyclic genes with common expression were defined using k-means clustering as described above. Cyclic genes with common function were defined as those which shared the same GO annotation. For the purpose of predicting cyclic expression, we used only those GO annotations over-enriched in at least one phase of cyclic expression and where at least 8 annotated genes were over-enriched in the same phase.

ACKNOWLEDGEMENTS

We thank Alexander Seddon and John Lloyd for help in identifying pCREs. We also thank all the members of the algae group at Michigan State University for their attention and critique. This work was supported by a Michigan State University Strategic Partnership Grant to C.B, E. F., and S.-H. S. and a National Science Foundation grant (MCB-1119778) to S.-H. S.

APPENDIX

Supplemental Materials and Methods

Determining threshold scores for COSPOT and DFT

Mittag et al. (2005) lists 18 proteins in *C. reinhardtii* which have previously show to exhibit circadian changes in the rate of transcription or concentration of mRNA. Amino acids sequences of these proteins were identified through KEGG and mapped to the *C. reinhardtii* genome using the TBLASTN tool available through Phytozome. We found fifteen proteins which mapped unambiguously to the *C. reinhardtii* genome and had matching annotation, only one of which was not present in the mRNA seq data set (**Supplemental Table 2.4**).

To define a cutoff threshold for each of our methods, COSPOT and the DFT, each program was evaluated against our gold-standard set three different *p*-values thresholds (or the equivalent cyclic-score): 0.05, 0.02, and 0.01. For each *p*-value threshold, the coverage of both the gold-standard set and the whole *C. reinhardtii* genome is reported in **Supplemental Table 2.5**. At each *p*-value threshold, the union of predictions from methods was used to define cyclically expressed genes in *C. reinhardtii*. As such, the *p*-value of the new two-dimensional threshold is defined by the joint distribution of COSPOT and DFT scores. Calculating this value is complicated by the fact that these scores are highly correlated ($r^2 > 0.7$), but the joint probability can be estimated using a randomized population of expression vectors (**Supplemental Table 2.6**). For every test *p*-value threshold, the increase in joint probability (compared to individual significant thresholds) was relatively moderate whereas coverage of the gold standard increased by as much 20% over a single method. We chose to use the combination of COSPOT&DFT as our predictive method with a test *p*-value threshold of 0.02, which balances in the inflation of the joint probability with the coverage of the gold standard set.

The combined method is most effective at excluding non-cycling genes, rather than defining cycling genes, which can be seen by looking at the correlation of both methods at different scoring threshold (**Supplemental Figure 2.7**). While the overall correlation between both methods is high, the correlation amongst highly scoring genes (exceeding the 0.02 threshold for either method) is actually quite low ($r^2 < 0.2$). Genes which score very highly with one method may be at or just below the margin for the other, however, a gene which scores poorly in one method generally scores poorly with the other. Therefore, we chose a more conservative score threshold as a cautionary measure.

Derivation of the model of duplicate gene divergence

Divergence of expression state was modeled using the following system of difference equations:

$$C_{t+1} = C_t \cdot (1-d) + \frac{D_t s}{2} \qquad (3)$$

$$N_{t+1} = N_t \cdot (1-d) + \frac{D_t s}{2} \qquad (4)$$

$$D_{t+1} = D_t \cdot (1-s) + C_t d + N_t d \qquad (5)$$

Where C, N, and D represent the frequencies of cycling, non-cycling, and divergent duplicates at a given *Ks* (subscript "t") and the subsequent *Ks* (subscript "t+1"). The variables d and s are, respectively, the probabilities of divergence from the identical state and reversion to the identical state. Since the null model assumes no bias, d and s are insensitive to whether the identical state is cycling or non-cycling.

Solving equations (3) and (4) for d, we obtain:

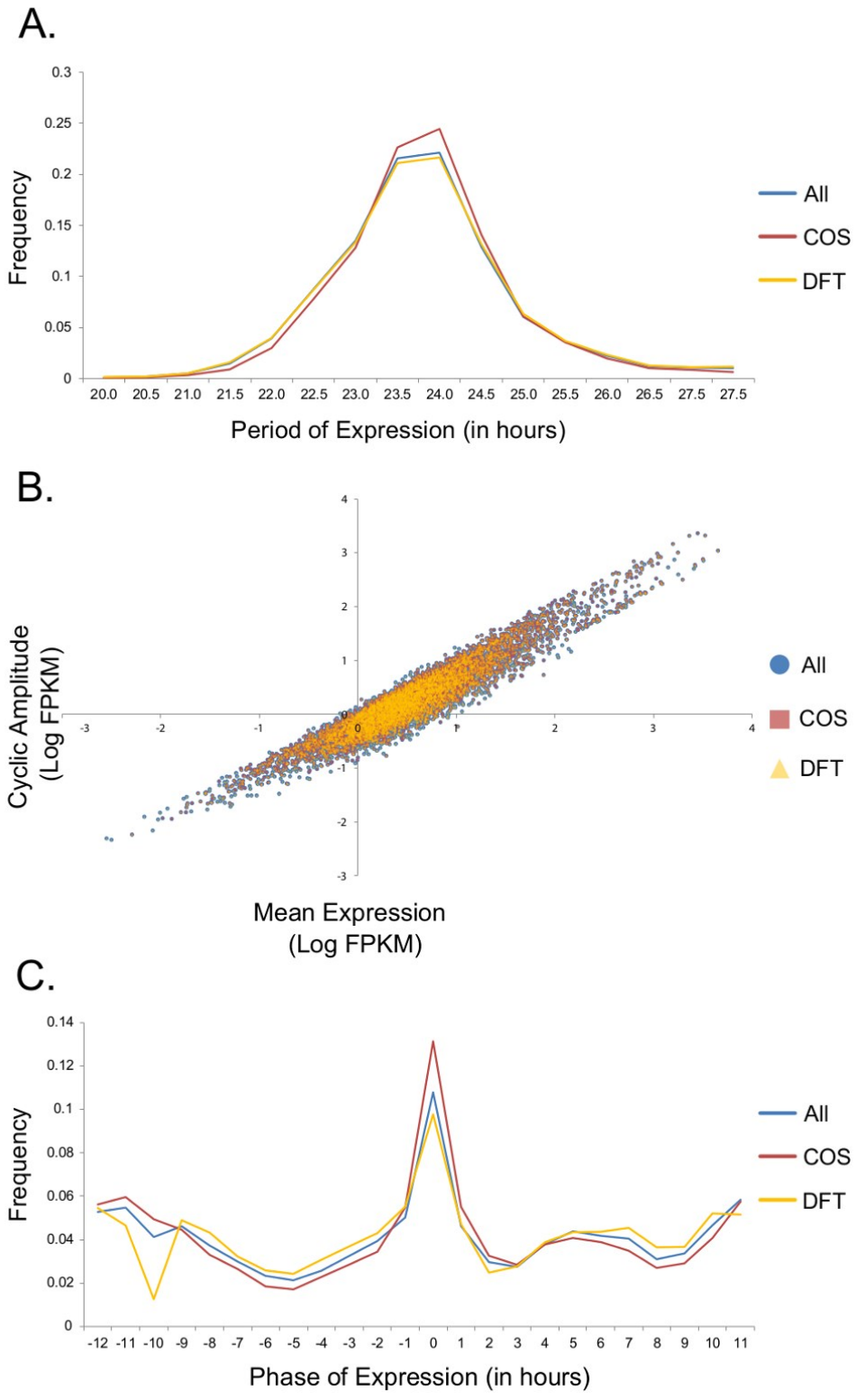
$$d = 1 + \frac{D_t s}{2C_t} - \frac{C_{t+1}}{C_t} \qquad (6)$$

$$d = 1 + \frac{D_t s}{2N_t} - \frac{N_{t+1}}{N_t} \qquad (7)$$

Using the property that the right hand sides of (6) and (7) must be equal, we arrive at the following formula for s that depends solely on duplicate frequencies:

$$s = \frac{\frac{2C_{t+1}}{C_t D_t} - \frac{2N_{t+1}}{N_t D_t}}{\frac{1}{C_t} - \frac{1}{N_t}} \qquad (8)$$

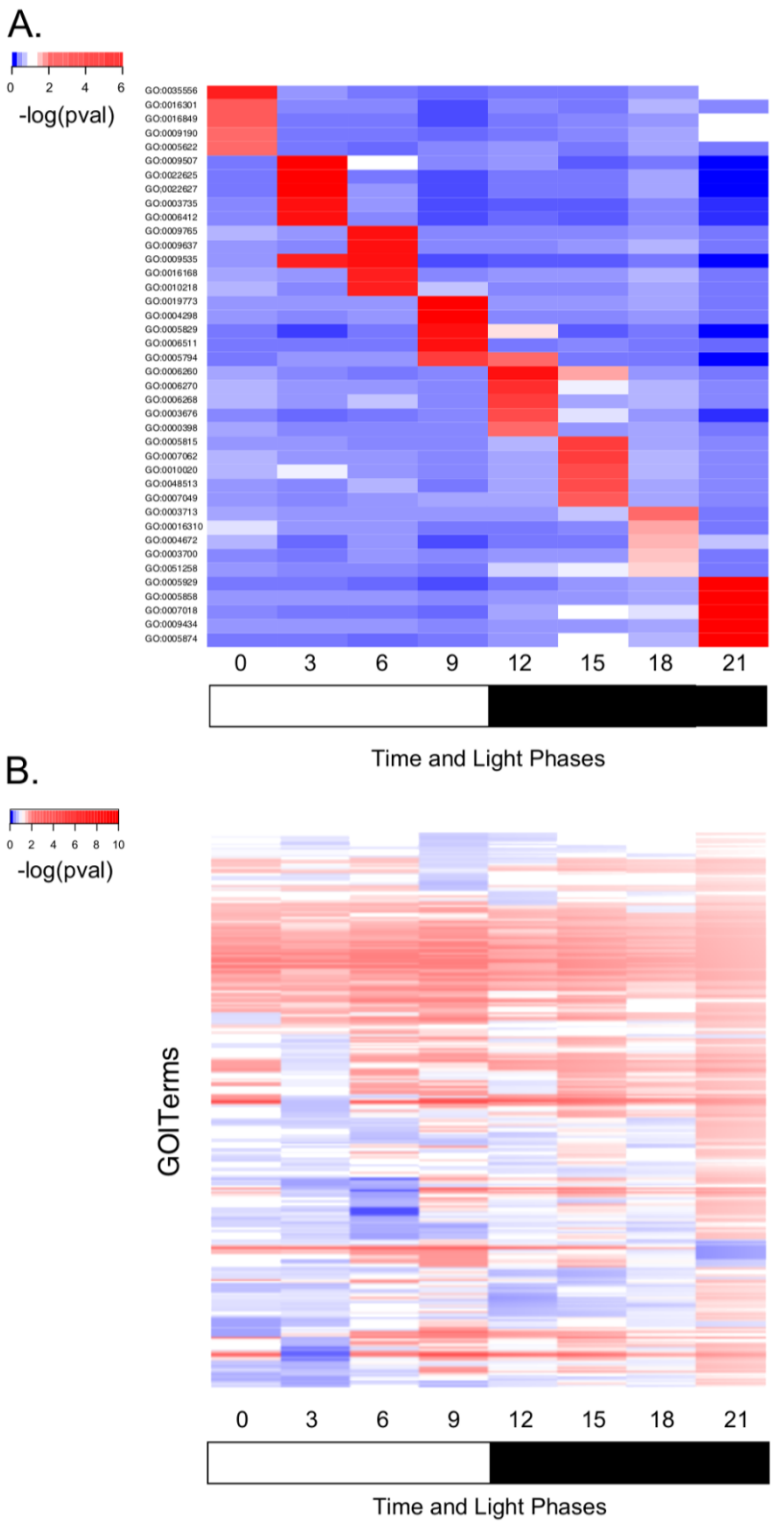
Initial conditions were set equal to values of C, D, and N observed at *Ks* = 0.3. We first attempted to fit values for d and s using frequencies at *Ks* 0.3 and 0.6, however because the percentage change in C is greater than N, we obtained a negative value for s. Since s is a probability, this results is unrealistic, so instead we fit d and s using *Ks* 0.6 and 0.9, obtaining values for d (0.42) and s (0.53) that were within [0,1]. Using these parameters, our model was able to replicate the overall behavior we observed, including the initial dip in C, though the percentage change is less than that of N (**Supplemental Figure 2.3**). The root mean squared error between our predictions and observation was 0.03.



Supplemental Figure 2.1 Period, amplitude, and phase of cyclic expression amongst predictions made by COSPOT, DFT, and both methods combined. (A) The distribution of

Supplemental Figure 2.1 (cont'd)

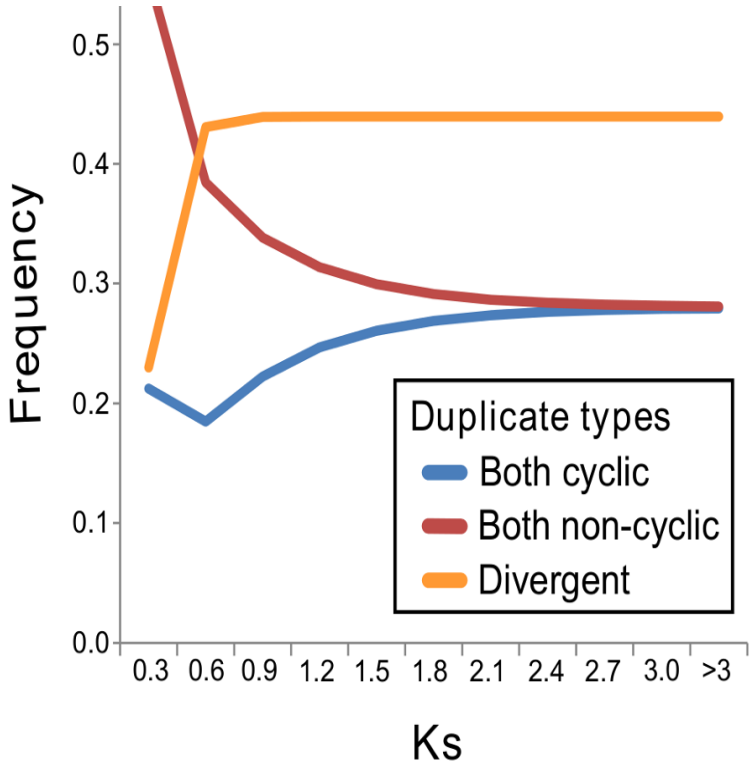
the period of expression in cycling genes predicted by COSPOT (red), DFT (yellow) and both methods combined (blue). (B) Mean expression (x-axis) vs. the amplitude of cyclic expression (y-axis) of cycling genes. Color labels follow (A). (C) Phase of expression of cycling genes. Color labels follow (A).



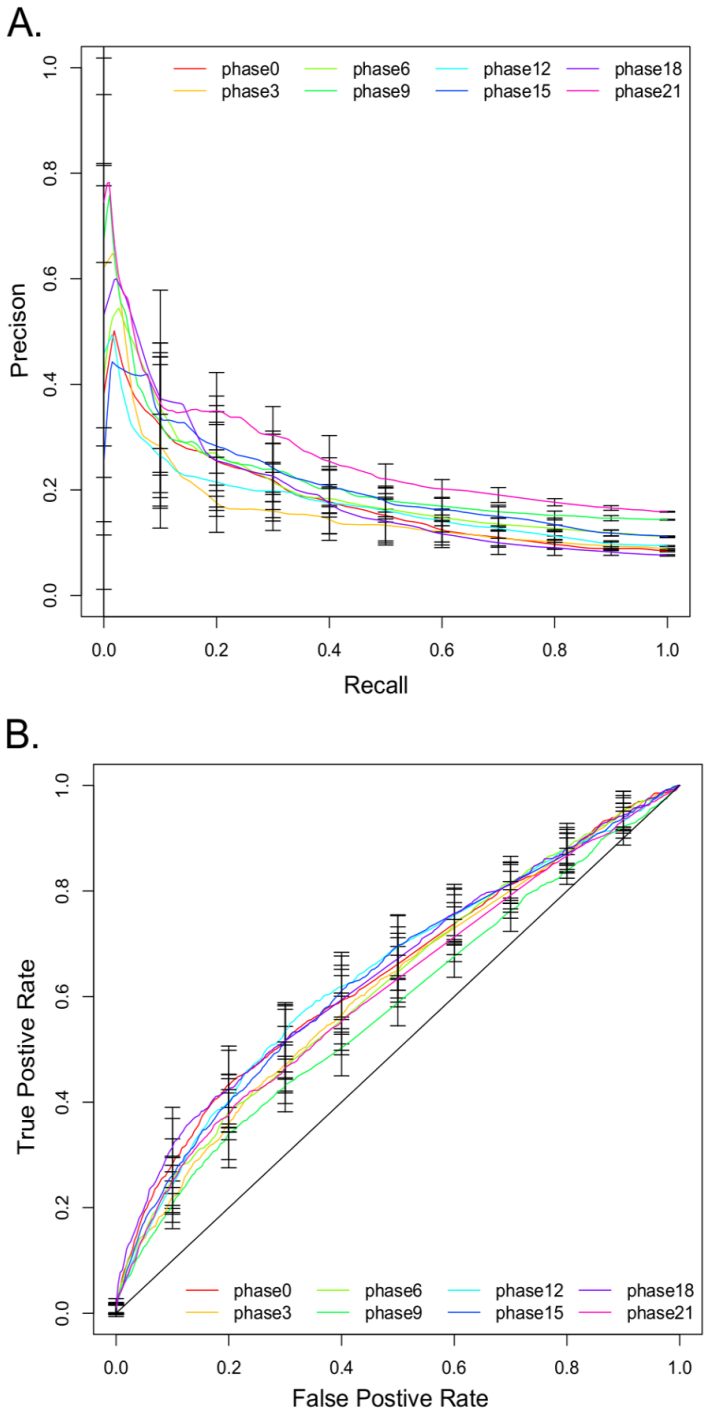
Supplemental Figure 2.2 Most over- and under-enriched GO terms amongst phase clusters of cycling genes. (A) Heatmap showing the –log10 transformed Fisher's exact test *p*-values (pval) of the top five GO terms with over-represented numbers of genes in each phase cluster

Supplemental Figure 2.2 (cont'd)

(ZT 0, 3, 6, 9, 12, 15, 18, and 21). (B) Heatmap showing transformed p-values of GO terms with under-represented numbers of genes in at least one phase cluster (same as in (A)). P-values were calculated and transformed as in part (A) except that the left-tail p-value was used.



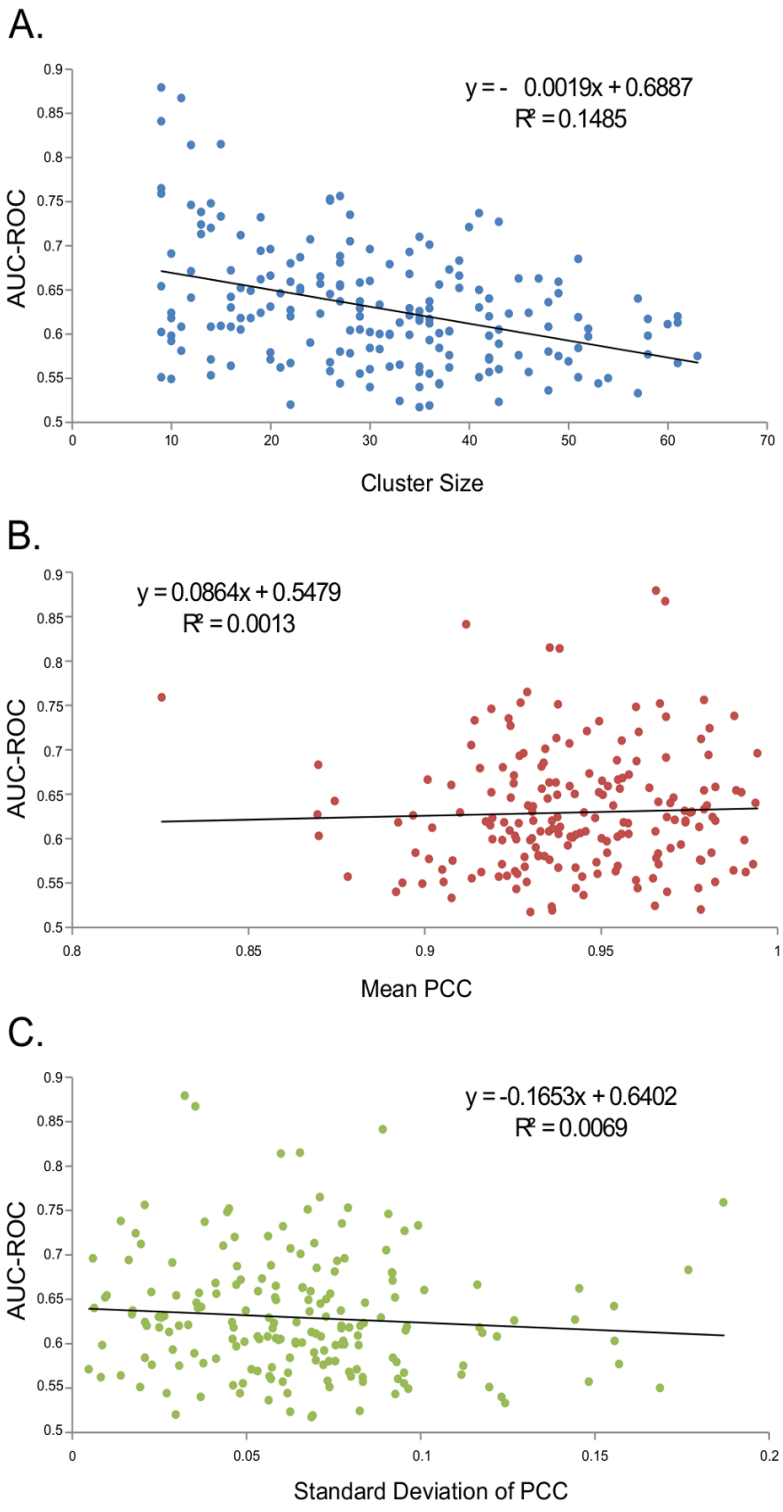
Supplemental Figure 2.3 Divergence of duplicate gene expression state modeled as a system of difference equations. The frequency at which duplicate pairs in *C. reinhardtii* are both cycling (blue), both non-cycling (red), or divergent expression (orange) as a function of the synonymous substitution rate (*Ks*). The difference equations used to generate these data are described in the Supporting Information.



Supplemental Figure 2.4 Precision-recall and AUC-ROC curves of SVM predictions for *C. reinhardtii*. (A) Precision-recall curves for the prediction of the each of the eight phase clusters in cycling genes in *C. reinhardtii* as classified using SVM. Each phase-cluster is represented as a different colored line: 0 (red), 3 (orange), 6 (lime), 9 (green), 12 (teal), 15 (blue), 18 (purple), 21 (pink). Error bars represent the variance in 10 separate runs of the SVM classifier at optimal

Supplemental Figure 2.4 (cont'd)

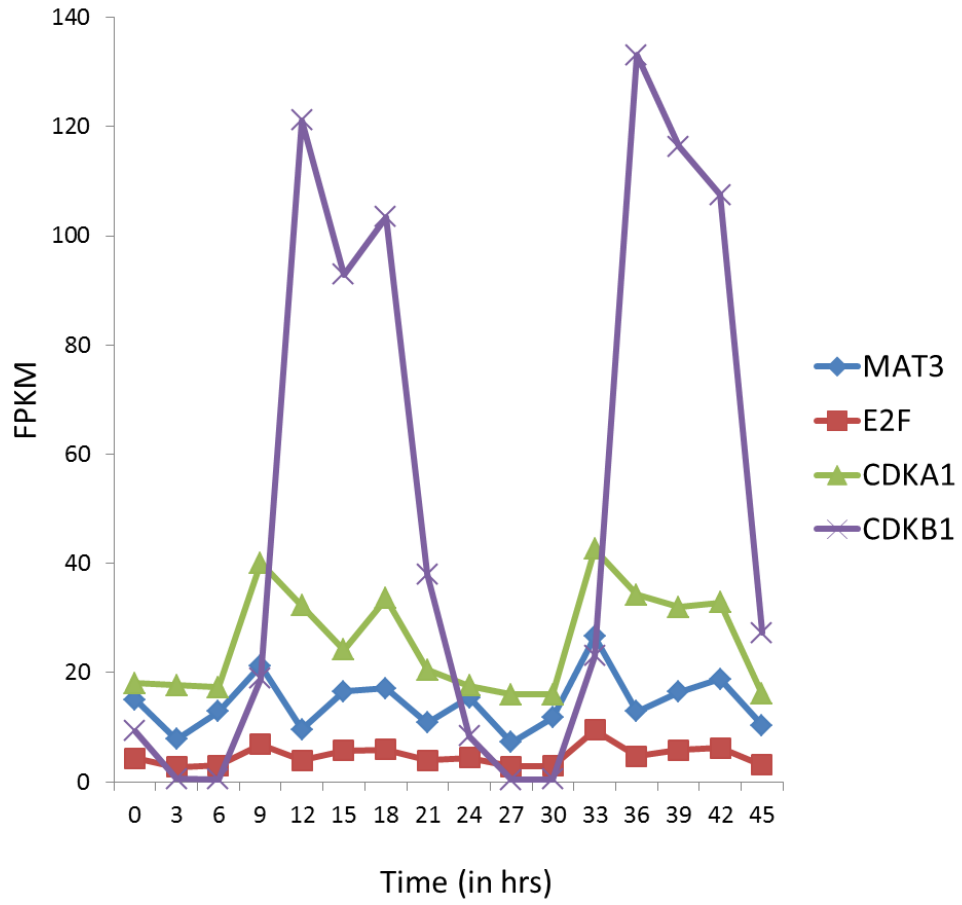
parameters. (B) ROC curves for the prediction of each of the eight phase clusters in cycling genes in *C. reinhardtii* as classified using SVM. Line color and error bars are assigned as in (A).



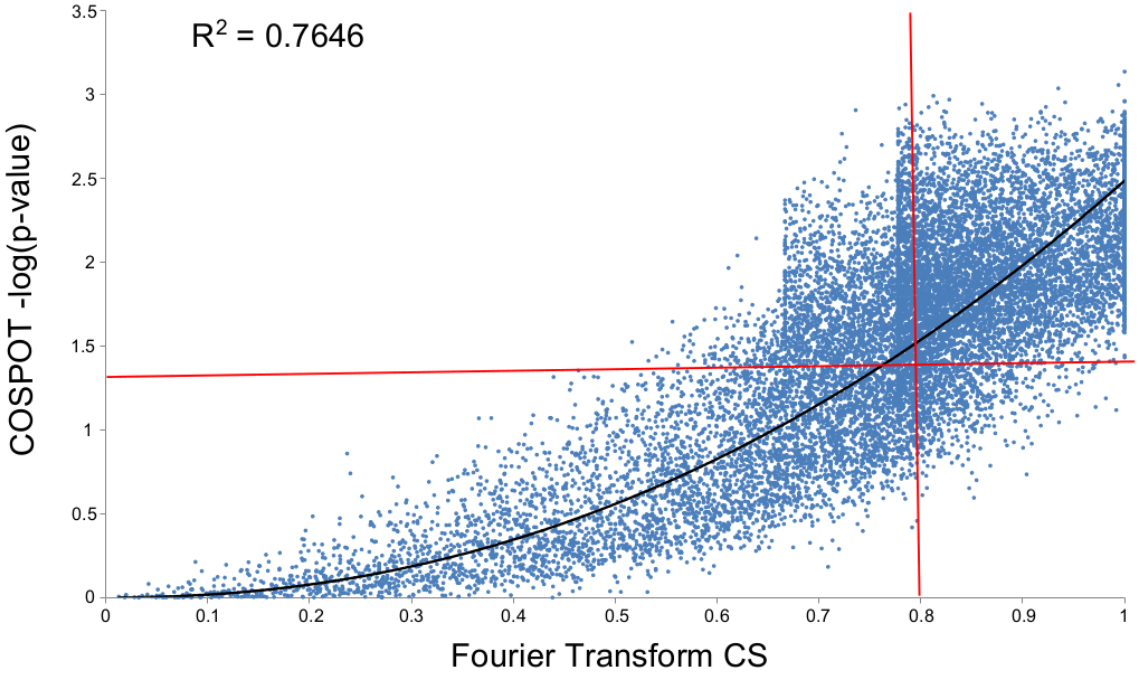
Supplemental Figure 2.5 Regression of the AUC-ROC of phase-expression clusters against cluster size, and Pearson Correlation Coefficient (PCC) of genes in the cluster. (A) Plot of phase-expression cluster size against AUC-ROC. The black line indicates the best linear

Supplemental Figure 2.5 (cont'd)

regression of AUC-ROC against cluster size. The equation is reported above the figure. (B) Plot of the mean PCC amongst genes in each phase-expression cluster against AUC-ROC. The black line indicates the best linear regression of AUC-ROC against mean PCC. The equation is reported above the figure. (C) Plot of the standard deviation of PCC amongst genes in each expression cluster against AUC-ROC. The black line indicates the best linear regression of AUC-ROC against standard deviation of PCC. The equation is reported above the figure.



Supplemental Figure 2.6 Expression profiles of cell cycle genes (MAT3, E2F, CDKA1, and CDKB1) in C. reinhardtii grown in TAP (Tris-Acetate-Phosphate) culture. As observed in previous studies of C. reinhardtii grown on autotrophic conditions (BISOVA et al. 2005), MAT3, CDKA1, and CDKB1 are most highly expressed between 12 and 18 hours after dawn, while E2F expression increases slightly earlier (between 6 and 9 hours).



Supplemental Figure 2.7 Distribution of Fourier Transform cyclic score and COSPOT p-values. Plot of Fourier Transform cyclic score (x-axis) against the negative log transform of the COSPOT *p*-value (y-axis). The black line is the best fit power-law regression of the transformed COSPOT *p*-value against Fourier Transform cyclic score. The red lines indicated the score threshold at a significance level of $\alpha < 0.02$ for the Fourier Transform cyclic score (vertical) and the transformed COSPOT *p*-value (horizontal).

Supplemental Table 2.1 Distribution of Fourier Transform cyclic score and COSPOT p-values

GO Term	adjusted p-value[1]	Description
GO:0019861	4.38E-28	flagellum
GO:0005929	3.07E-11	cilium
GO:0035086	4.41E-06	axoneme
GO:0005874	1.16E-04	microtubule
GO:0007018	4.59E-04	microtubule-based movement
GO:0010287	4.59E-04	regulation of glucose transport
GO:0003777	6.53E-04	microtubule motor activity
GO:0009765	1.19E-03	carbohydrate mediated signaling
GO:0022625	1.19E-03	cytosolic large ribosomal subunit
GO:0006260	1.71E-03	DNA replication
GO:0030286	1.89E-03	dynein complex
GO:0005886	2.16E-03	plasma membrane
GO:0030030	6.12E-03	cell projection organization
GO:0005794	6.52E-03	Golgi apparatus
GO:0042995	1.00E-02	cell projection
GO:0005198	1.00E-02	structural molecule activity
GO:0003774	1.00E-02	motor activity
GO:0022627	1.00E-02	cytosolic small ribosomal subunit
GO:0005774	1.42E-02	vacuolar membrane
GO:0009653	1.77E-02	anatomical structure morphogenesis
GO:0009535	1.77E-02	chloroplast thylakoid membrane
GO:0009637	2.20E-02	response to blue light
GO:0006364	2.20E-02	rRNA processing
GO:0005932	2.65E-02	microtubule basal body
GO:0009506	2.65E-02	plasmodesmata
GO:0004674	2.65E-02	protein serine/threonine kinase

Supplemental Table 2.1 (cont'd)

GO:0005509	2.66E-02	calcium ion binding
GO:0005488	2.75E-02	binding
GO:0030992	2.75E-02	intraciliary transport particle B
GO:0009507	3.12E-02	chloroplast
GO:0010114	3.34E-02	response to red light
GO:0010218	3.34E-02	response to far red light
GO:0046686	3.79E-02	response to cadmium ion
GO:0009523	4.16E-02	photosystem II
GO:0048046	4.21E-02	apoplast
GO:0006270	4.21E-02	DNA replication initiation
GO:0009296	4.21E-02	flagellum assembly
GO:0010020	4.21E-02	chloroplast fission
GO:0009434	4.21E-02	motile cilium
GO:0044430	4.21E-02	cytoskeletal part
GO:0019253	4.21E-02	reductive pentose-phosphate cycle
GO:0019773	4.21E-02	proteasome core complex, alpha-subunit complex
GO:0009826	4.21E-02	unidimensional cell growth
GO:0004298	4.33E-02	threonine-type endopeptidase activity

1. Fisher Exact Test *p*-value adjusted according to Benjamini-Hochberg

Supplemental Table 2.2 Descriptions of the GO terms in each of the five broad functional categories

Category	GO Terms	Description
photosynthesis and light response	GO:0015671	oxygen transport
	GO:0009773	photosynthetic electron transport in photosystem I
	GO:0010206	photosystem II repair
	GO:0009765	photosynthesis, light harvesting
	GO:0015979	photosynthesis
	GO:0010218	response to far red light
	GO:0009637	response to blue light
	GO:0010114	response to red light
	GO:0010304	PSII associated light-harvesting complex II catabolic process
	GO:0010020	chloroplast fission
	GO:0009507	chloroplast
	GO:0009579	thylakoid
	GO:0009570	chloroplast stroma
	GO:0009941	chloroplast envelope
	GO:0009543	chloroplast thylakoid lumen
	GO:0009523	photosystem II
	GO:0009522	photosystem I
	GO:0009533	chloroplast stromal thylakoid
	GO:0009534	chloroplast thylakoid
	GO:0010287	plastoglobule
	GO:0009535	chloroplast thylakoid membrane
	GO:0016168	chlorophyll binding
cell cycle and mitosis	GO:0006260	DNA replication
	GO:0006270	DNA replication initiation
	GO:0006268	DNA unwinding involved in replication
	GO:0000910	cytokinesis
	GO:0000724	double-strand break repair via homologous recombination
	GO:0006302	double-strand break repair
	GO:0007062	sister chromatid cohesion
	GO:0007067	mitosis
	GO:0006259	DNA metabolic process
	GO:0007049	cell cycle
	GO:0051726	regulation of cell cycle

Supplemental Table 2.2 (cont'd)

	GO:0006281	DNA repair
	GO:0051301	cell division
	GO:0006310	DNA recombination
	GO:0005819	spindle
	GO:0005815	microtubule organizing center
	GO:0005694	chromosome
	GO:0004003	ATP-dependent DNA helicase activity
	GO:0003887	DNA-directed DNA polymerase activity
	GO:0004386	helicase activity
microtubules and	GO:0000226	microtubule cytoskeleton organization
flagella	GO:0007018	microtubule-based movement
	GO:0009296	flagellum assembly
	GO:0030030	cell projection organization
	GO:0042384	cilium assembly
	GO:0015630	microtubule cytoskeleton
	GO:0044430	cytoskeletal part
	GO:0019861	flagellum
	GO:0030286	dynein complex
	GO:0035086	cilium axoneme
	GO:0005813	centrosome
	GO:0005932	microtubule basal body
	GO:0005874	microtubule
	GO:0005929	cilium
	GO:0005856	cytoskeleton
	GO:0005858	axonemal dynein complex
	GO:0035085	cilium axoneme
	GO:0009434	motile cilium
	GO:0030992	intraflagellar transport particle B
	GO:0042995	cell projection
	GO:0044463	cell projection part
	GO:0005876	spindle microtubule
	GO:0003777	microtubule motor activity
	GO:0003774	motor activity
	GO:0004835	tubulin-tyrosine ligase activity

Supplemental Table 2.2 (cont'd)

mitochondria and	GO:0006096	glycolysis
metabolism	GO:0006122	mitochondrial electron transport, ubiquinol to cytochrome c
	GO:0005983	starch catabolic process
	GO:0006098	pentose-phosphate shunt
	GO:0007005	mitochondrion organization
	GO:0006508	proteolysis
	GO:0015986	ATP synthesis coupled proton transport
	GO:0045261	proton-transporting ATP synthase complex, catalytic coreF(1)
	GO:0005750	mitochondrial respiratory chain complex III
	GO:0005747	mitochondrial respiratory chain complex I
	GO:0005739	mitochondrion
	GO:0005759	mitochondrial matrix
	GO:0005741	mitochondrial outer membrane
	GO:0005743	mitochondrial inner membrane
	GO:0046933	proton-transporting ATP synthase activity, rotational
	GO:0046961	mechanism
		proton-transporting ATPase activity, rotational mechanism
ribosome and	GO:0006414	translational elongation
translation	GO:0006412	translation
	GO:0022626	cytosolic ribosome
	GO:0022625	cytosolic large ribosomal subunit
	GO:0022627	cytosolic small ribosomal subunit
	GO:0019843	rRNA binding
	GO:0003735	structural constituent of ribosome

Supplemental Table 2.3: Optimal parameters and performance measures of SVM classification

Phase	C[1]	R[2]	AUC-ROC	F-measure	Precision	Recall
0	0.01	4	0.64	0.22	0.24	0.20
3	0.1	1.5	0.62	0.21	0.14	0.39
6	0.1	4	0.62	0.19	0.26	0.15
9	0.1	2.5	0.58	0.22	0.27	0.18
12	0.01	3.5	0.64	0.19	0.21	0.18
15	0.1	1.5	0.64	0.27	0.21	0.40
18	0.1	4	0.65	0.23	0.24	0.23
21	0.01	3.5	0.61	0.21	0.38	0.15

1. C = minimum separation

2. R = ratio of negative to positive examples

Supplemental Table 2.4 "Gold Standard" cycling genes in *C. reinhardtii*

Gene	Name	Reference	Locus	DFT Cyclic Score	COSPOT p-value
ATP2/ARF1	ADP-ribosylation factor	MEMON *et al.* (1995)	Cre17.g698000	0.90	1.9e-02
CAH1	carbonic anhydrase	FUJIWARA *et al.* (1996)	Cre04.g223100	0.76	1.7e-01
CYC4	cytochrome c	JACOBSHAGEN *et al.* (2001)	Cre16.g670950	0.78	6.2e-01
Cytosolic thioredoxin h1	Cytosolic thioredoxin h1	LEMAIRE *et al.* (1999)	Cre09.g391900	0.29	4.0e-01
FBA1	chloroplastic fructose-bisphosphate aldolase	JACOBSHAGEN *et al.* (2001)	Cre01.g006950	0.95	1.2e-02
FBA2	chloroplastic fructose-bisphosphate aldolase	JACOBSHAGEN *et al.* (2001)	Cre02.g093450	0.83	3.4e-02
FBA3	chloroplastic fructose-bisphosphate aldolase	JACOBSHAGEN *et al.* (2001)	Cre05.g234550	0.9	2.3e-02
FBA4	chloroplastic fructose-bisphosphate aldolase	JACOBSHAGEN *et al.* (2001)	Cre02.g115650	0.61	1.9e-02

Supplemental Tabel 2.4 (cont'd)

FNR1	Ferredoxin NADP reductase	LEMAIRE *et al.* (1999)	Cre11.g476750	0.81	2.4e-02
HSP70B	70kd family heat shock protein	JACOBSHAGEN *et al.* (2001)	Cre06.250100	0.32	8.5e-01
LCHII	Chlorophyll binding protein	JACOBSHAGEN *et al.* (1996)	Cre06.g283950	0.82	6.5e-03
PRK1	phosphoribulokinase	LEMAIRE *et al.* (1999)	Cre12.g554800	0.99	1.0e-02
TUB1	Beta-tubulin	JACOBSHAGEN & JOHNSON (1994)	Cre12.g542250	0.67	1.1e-02
TUB2	Beta-tubulin	JACOBSHAGEN & JOHNSON (1994)	Cre12.g549550	0.98	1.1e-02
TufA	Elongation factor Tu	HWANG *et al.* (1996)	Cre06.g259150	0.77	1.1e-02

Supplemental Table 2.5 Performance COSPOT and DFT on *C. reinhardtii*

Method and α levels	Genome Coverage[1]	Gold Stand Coverage[2]
COSPOT		
α = 0.01	21.0% (3590)	6.7% (1)
α = 0.02	37.4% (6400)	46.7% (7)
α = 0.05	54.9% (9392)	73.3% (11)
DFT		
α = 0.01	29.6% (5061)	33.3% (5)
α = 0.02	37.6% (6443)	53.3% (8)
α = 0.05	55.8% (9556)	73.3% (11)

1. Parentheses indicated the actual number of genes covered

2. Parentheses indicated how many of 15 gold standard genes are identified as cyclic

Supplemental Table 2.6 Performance of combining COPSOT and DFT on *C. reinhardtii*

Test P-value	Joint Probability[1]	Overlap[2]	Genome Coverage[3]	Gold Stand Coverage[4]
α = 0.01	0.0134	39.6% (2414)	35.7% (6236)	40.0% (6)
α = 0.02	0.0272	56.7% (4579)	47.2% (8072)	73.3% (11)
α = 0.05	0.0734	73.5% (8024)	61.7% (10552)	86.6% (13)

1. The joint probability of a gene having a score with a p-value of α in either COSPOT or DFT

2. Parentheses indicated the actual number of genes in the overlap set

3. Parentheses indicated the actual number of genes covered

4. Parentheses indicated how many of 15 gold standard genes are identified as cyclic

Supplemental File 2.1

Supplemental File 2.1 can be found at the following link:

http://www.g3journal.org/highwire/filestream/472423/field_highwire_adjunct_files/10/FileS3.xlsx

Supplemental File 2.2

Supplemental File 2.2 can be found at the following link:

http://www.g3journal.org/highwire/filestream/472423/field_highwire_adjunct_files/11/FileS4.xlsx

Supplemental File 2.3

Supplemental File 2.3 can be found at the following link:

http://www.g3journal.org/highwire/filestream/472423/field_highwire_adjunct_files/12/FileS5.txt

Supplemental File 2.4

Supplemental File 2.4 can be found at the following link:

http://www.g3journal.org/highwire/filestream/472423/field_highwire_adjunct_files/13/FileS6.xlsx

Supplemental File 2.5

Supplemental File 2.5 can be found at the following link:

http://www.g3journal.org/highwire/filestream/472423/field_highwire_adjunct_files/14/FileS7.txt

Supplemental File 2.6

Supplemental File 2.6 can be found at the following link:

http://www.g3journal.org/highwire/filestream/472423/field_highwire_adjunct_files/15/FileS8.xlsx

Supplemental File 2.7

Supplemental File 2.7 can be found at the following link:

http://www.g3journal.org/highwire/filestream/472423/field_highwire_adjunct_files/16/FileS9.txt

Supplemental File 2.8

Supplemental File 2.8 can be found at the following link:

http://www.g3journal.org/highwire/filestream/472423/field_highwire_adjunct_files/9/FileS2.zip

CHAPTER 3: PREDICTING CELL-CYCLE EXPRESSED GENES IDENTIFIES CANONICAL AND NON-CANONICAL REGULATORS OF TIME-SPECIFIC EXPRESSION IN *SACCHAROMYCES CEREVISIAE*

ABSTRACT

Gene expression is controlled by regulatory proteins know as transcription factors (TFs). The collection of all TFs, target genes and their interactions in an organism form a gene regulatory network (GRN), which can produce complex patterns of expression, such as cycling. However, identifying which interactions regulate expression in a specific context remains a challenging task complicated by the existence of multiple approaches to characterize GRNs. To assess how different methods of defining GRNs capture their regulatory function, we predicted general and phase-specific cell-cycle expression in *Saccharomyces cerevisiae* using four regulatory data sets: chromatin immunoprecipitation (ChIP), TF deletion data (Deletion), protein binding microarrays (PBMs), and position weight matrices (PWMs). Our results indicate that data sets with the highest coverage of the *S. cerevisiae* GRN (ChIP, Deletion and all PWMs) perform best in predicting cell-cycle expression. Furthermore, prediction performance was improved by including using TF-TF interactions from feed-forward loops as features as well as by combining the best predictive features from ChIP and Deletion data. The TFs that were the best predictors of cell-cycle expression were enriched for known cell-cycle regulators, but TFs important for predictive models built on ChIP and Deletion data were also enriched for GO annotations related to invasive growth and metabolism, respectively. Finally, analysis of important TF-TF interactions suggests that the GRN regulating cell cycle expression is highly interconnected and clustered around four groups of genes, two of which contain known cell-cycle regulators, while the other two contain TFs not previously identified as being involved in cell-cycle expression.

INTRODUCTION

Essential biological processes, from the replication of single cells (Spellman et al. 1998) to the development of multicellular organisms (Tomancak et al. 2002), are dependent on complex spatially and temporally specific patterns of gene expression. The regulation of gene expression during the initiation of transcription depends on both core promoter elements that interact with the RNA-Polymerase II complex (Juven-Gershon et al. 2008) and accessory elements known as transcription factors (TFs) that further promote or block the recruitment of RNA-Polymerase to particular promoter regions (Lelli, Slattery, and Mann 2012; Spitz and Furlong 2012). TFs bind to short DNA-sequences called *cis*-regulatory sites, often located in the upstream promoter region of a gene, though not all of these sites are necessarily occupied at all times (Lelli, Slattery, and Mann 2012; Spitz and Furlong 2012). TFs do not work in isolation to regulate gene expression. For example changes in the chromatin state around a promoter can impact TF binding (M. Li et al. 2015; Benveniste et al. 2014; Miller and Widom 2003). TFs also interact with other TFs. These interactions can be direct, such as cooperative binding of regulatory sites (Jolma et al. 2015; Kazemian et al. 2013) or indirect, such as collaborative/competitive binding to sites (Miller and Widom 2003). There can also be higher-order regulation, where the expression of one TF is regulated by other TFs, such that expression a gene may depend directly on the TF binding to its promoter and indirectly on the regulation of that TF. The sum total of direct (TF-target gene) and higher order (TF-TF) interactions regulating transcription in an organism is referred to as a gene regulatory network or GRN (Macneil and Walhout 2011). However, because TF binding is dependent on cooperative binding, cofactors, the chromatin state, and the abundance of the TF under the current conditions (Spitz and Furlong 2012), these direct and higher order interactions are not static. Therefore,

when attempting to understand complex patterns of gene expression, it is important to identify the relevant interactions.

Although it is understood that the interaction between a TF and the promoter of a target gene is mediated by *cis*-regulatory elements, inferring whether a TF binds to a specific promoter *in vivo* is complicated by the fact that TFs may recognize multiple distinct nucleotides sequences (Badis et al. 2009). The promiscuity of TF binding can be addressed by representing the size of the binding motif and nucleotide preference at different positions of the motif as a position weight matrix (PWM, Y. Li et al. 2015; Wasserman and Sandelin 2004; Stormo et al. 1982). PWMs of putative *cis*-regulatory elements can be identified without experimental evidence of TF binding by looking for the overrepresentation of DNA sequences in the promoters of coregulated genes using computational models (Wasserman and Sandelin 2004; Y. Li et al. 2015). Alternatively, the affinity between TFs and their binding sequence(s) can be assayed *in vitro* using protein binding microarrays (Bulyk 2007; Berger and Bulyk 2009) or *in vivo* by chromatin immunoprecipitation (ChIP, Buck and Lieb 2004; Furey 2012) or with other emerging technologies like DapSeq (O'Malley et al. 2016). Binding site information from these *in vivo*/*in vitro* assays can be used to define TF-specific PWMs (de Boer and Hughes 2012), or regulatory interactions can be identified directly by mapping binding sequences/reads to an annotated genome (Bailey et al. 2013). Finally, regulatory interactions can be identified by screening for differentially expressed genes in TF knockouts (Reimand et al. 2010). As there is no single characterization of a regulatory interaction, it is important to have a method to assess how well a GRN explains a specific expression pattern.

Before we use regulatory interactions to explain complex expression patterns, we first must define what we mean by a pattern of expression. Most simply, a gene's pattern of

expression is defined based on the magnitude of expression under a defined set of circumstances, such as a chosen environment/stress (Zou et al. 2011; Uygun et al. 2017), tissue/body part (Segal et al. 2008; Chikina et al. 2009), function/biological process (Beer and Tavazoie 2004; Panchy et al. 2014) or combination thereof (Uygun et al. 2017). Grouping genes based on clear criteria allows expression regulation to be approached as a classification problem where the presence or absence of a regulatory interaction is used to predict whether or not a gene exhibits the particular pattern of expression using a machine learning algorithm. To date, such machine learning algorithms have been applied to predicting expression in a variety of species, including *Caenorhabditis elegans* (Chikina et al. 2009), *Arabidopsis thaliana* (Zou et al. 2011; Uygun et al. 2017), and *Chlamydomonas reinhardtii* (Panchy et al. 2014)). This approach has been applied successfully to predict complex patterns of expression, such as tissue-specific response to stress (Uygun et al. 2017), while other modeling methods have been applied to co-expression clusters based on >100 different environmental conditions (Beer and Tavazoie 2004). Despite these successes, certain types of patterns remain challenging to predict, such as the specific timing of expression within a cyclic process (Panchy et al. 2014). Previous studies have explored improving the performance of machine learning-based predictions of gene expression by supplementing TF-interaction information with additional information such as DNA accessibility and cross-species conservation (Uygun et al. 2017) and by including information about combinatorial rules between TFs that can bind to the same promoter (Zou et al. 2011). However, it remains to be seen if these approaches are useful for predicting timing of gene expression or identifying regulatory interactions important for controlling that timing.

The cell cycle of budding yeast, *Saccharomyces cerevisiae,* (reviewed in Bahler 2005) is an ideal system for studying the regulation of complex expression patterns because the progression of this process is divided into distinct phases: initial growth (G1-phase), DNA replication (S-phase), intermediate growth (G2-phase), and cell-division (M-phase). Therefore, clear patterns of expression can be defined based on when genes reach their peak expression during the cell cycle, and the expression of genes during the cell cycle has been extensively characterized in *S. cerevisiae* (Price et al. 1991; Spellman et al. 1998). Furthermore, transcriptional regulation is known to play a key role in the control of cyclic expression during the cell cycle (Futcher 2002, Breeden 2003), and there are multiple data sets defining TF-target interactions in *S. cerevisiae* on a genome-wide scale (Harbison et al. 2004, Zhu et al. 2009, Reimand et al. 2010, de Boer and Hughes 2012). For these reasons, we used the *S. cerevisiae* cell-cycle as a model of complex expression in order to study the effect of different approaches for defining the yeast GRN on our ability to correctly characterize transcriptional regulation. The TF-target interactions were defined using PWMs, PBMs, ChIP-Chip, or Deletion data, and for each type of interaction data predictions were made using the same machine learning algorithms across all cell cycle time points, allowing us to examine the usefulness of each type of data. We also investigated whether performance could be improved by including TF-TF interactions as model inputs, applying feature selection algorithms to remove uninformative features, and by combining TF-interaction information from different data types. Once the best performing model was identified, Gene Ontology (GO) analysis was used to identify biological functions that are over- or under-represented in TFs most important for predicting the timing of cell cycle expression. Finally, we used the most important TF-TF interactions from our models to construct

putative GRNs, allowing us to identify subclusters of TFs whose interactions are central to controlling the timing of expression.

RESULTS AND DISCUSSION

Comparing TF-target interactions from multiple regulatory data sets

Although there is a single GRN which describes transcriptional regulation in an organism, different approaches to defining regulatory interactions may results in inferred GRNs that have greater or lesser degrees of similarity. In this study, TF-target interactions in *S. cerevisiae* was defined using four distinct data sets: TF binding sites inferred from ChIP-Chip experiments (ChIP-Chip), interactions inferred from changes in expression in deletion mutants (Deletion), TF binding sites inferred from PWMs (PWM), and TF binding sites inferred from PBM data (PBM) (**Table 3.1**). Further details about the processing of each data set can be found in the Methods section. Because of methodological differences, we would expect to find differences between GRNs defined using different data sets, both in the total number of regulatory interactions and the specific relationships between TFs and target genes. The number of TF-target interactions in the *S. cerevisiae* GRNs varies from 16,602 in the ChIP-Chip data set to 78,095 in the PWM data set. This almost 5-fold difference in the number of interactions identified is not due to differences in the amount of data; each data set includes at least 80 TFs and 4,701 annotated gene ORFs (**Table 3.1**). The large difference is driven instead by differences in the average number of interactions per TF, which varies from 105.6 in the ChIP-Chip GRN to 558.8 in the PBM GRN. The distribution of TF number per target is positively skewed for the ChIP-Chip (2.23), Deletion (4.04), and PWM (3.43) GRNs, indicating that most TFs have fewer interactions than the average value while a few have many more interactions. The majority of TFs were present in more than one data set (**Figure 3.1A**); however, the number of interactions that each TF is involved in is only weakly correlated between the ChIP and

Table 3.1 Size and origin of GRNs defined from each data set

Data Set	Transcription Factors	All Genes	Interactions	Source
ChIP-Chip	152	4701	16,062	ScerTF
Deletion	151	5256	26,757	ScerTF
PWM	230	6536	78,095	YeTFaSCO
Expert PWM	104	4740	9726	YeTFaSCO
PBM	81	4922	45,264	Zhu et al. (2009)

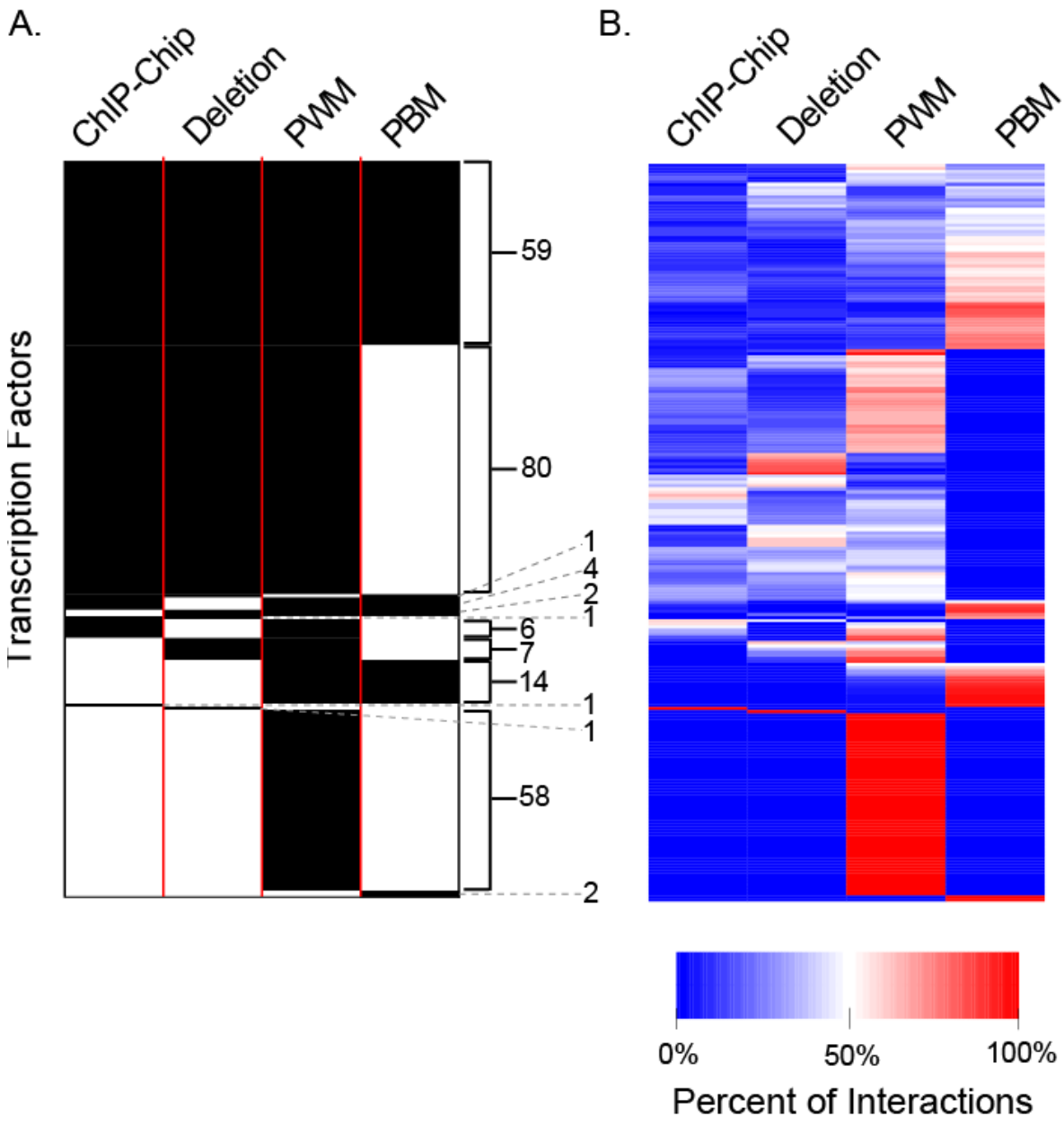


Figure 3.1 Coverage of TF and TF-interactions by data set (A) Heatmap of the coverage of *S. cerevisiae* TFs in GRNs derived from different data sets. Each row represents a TF and each column represents the GRN derived from a different data set (ChIP-Chip, Deletion, PWM, PBM). TFs are sorted according the GRNs they are found in such that TFs belonging to the same set of GRNs are grouped together. The number of TFs belonging to each group is indicated on

Figure 3.1 (cont'd)

the right side of the graph. (B) Heatmap of the percentage of TF-target interactions for each *S. cerevisiae* TF belonging to each GRN. Each row represents a TF, and each column represents the GRN derived from a different data set (ChIP-Chip, Deletion, PWM, PBM). Dark red indicates a higher percentage of interactions found within a data set, while dark blue indicates a lower percentage of interactions. TFs are ordered as in (A).

Deletion (Pearson's product moment correlation coefficient (PCC) = 0.092), ChIP and PWM (PCC = 0.109), and Deletion and PWM (PCC=0.046) datasets.

To further investigate the consistency of inferred TF-target interactions, for each TF we calculated the percentage of total interactions originating from each data set and grouped them using hierarchical clustering (**Figure 3.1B**). Although most TFs are found in > 1 GRN, TFs are primarily clustered based on the GRN in which they are most prominent, which is not unexpected given that for the majority (80.5%) of TFs, more than half of their interactions were identified from a single data type. This pattern held true even when TFs unique to a single data set were excluded: for 73.6% of TFs found in >1 GRN, more than half of their interactions were from a single data set. We also looked at the overlap of specific interactions (i.e. the same TF and target gene) between the different data sets, including a subset of the PWM data set including only curated binding sites (**Figure 3.2**). Of the 156,710 TF-target interactions identified, 89.0% were unique to a single data set, with 40.0% of unique interactions belonging to the PWM data set. As expected, there was a large overlap between the full PWM data set and the curated PWM subset, totaling 9,458 interactions or 96.8% of all interactions from the curated PWMs. However, the degree of overlap between the four main GRNs varied; when TF targets were chosen at random, ChIP-Chip overlap with Deletion (pv=2.37e-65) and PWM (pv<1e-307) was higher than expected by random chance, but PWM overlap with Deletion (pv=1.74e-111) and PBM (pv=1.87e-106) lower (see Methods). The number of overlaps between ChIP-Chip and PBM (0.057) and Deletion and PBM (0.43) were not significant in either direction (**Supplementary Figure 3.1**). This suggests that the ChIP-Chip data set is generally more similar to the other data sets, while PWMs are more dissimilar. Although, given the low overall

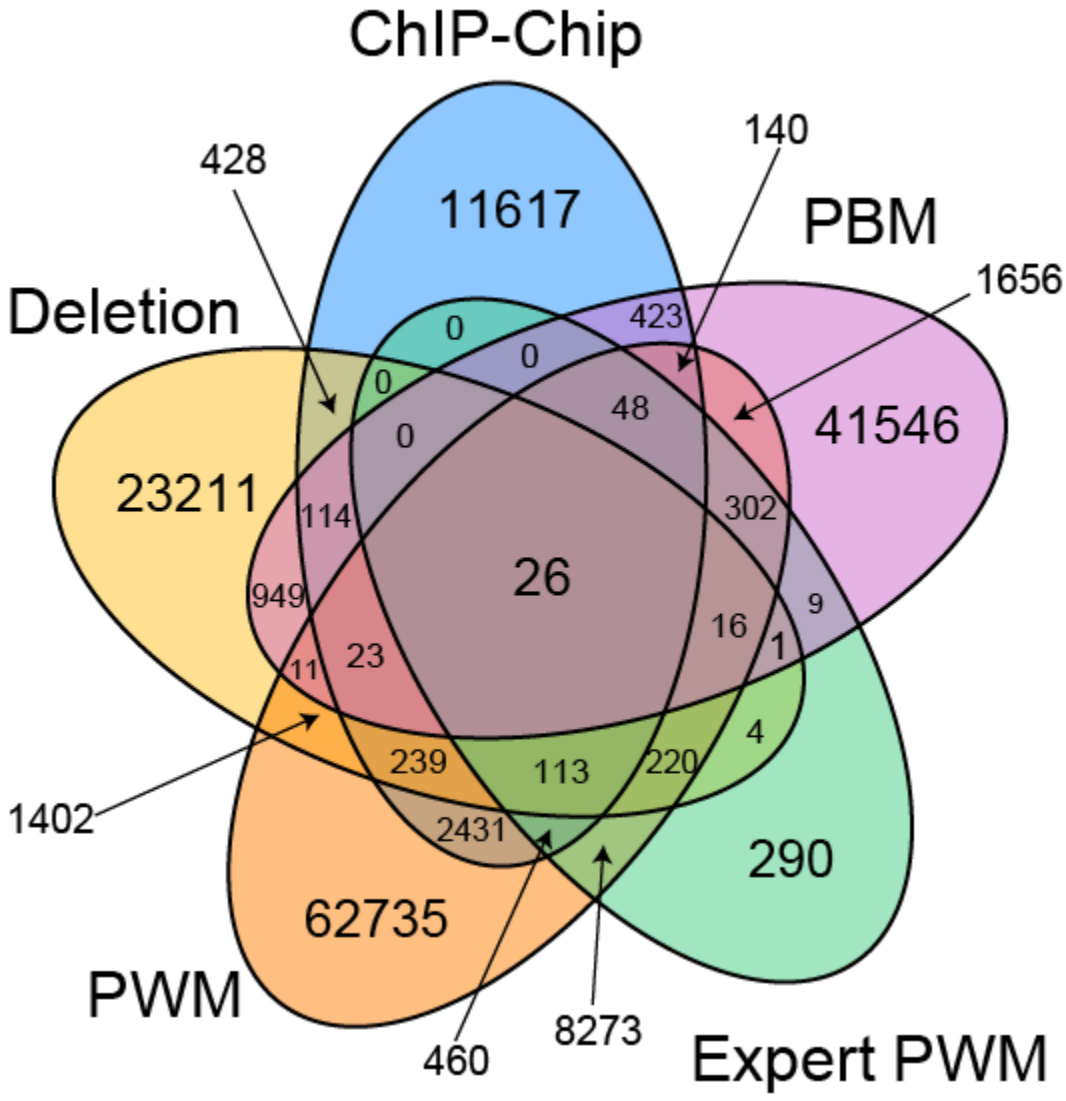


Figure 3.2 Overlap in TF-target interactions across data sets. Colors of different regions indicate different data sets: ChIP-Chip (blue), Deletion (yellow), PWM (orange), Expert PWM (green), PBM (purple).

overlap between data sets, we would still expect models built on each data set to perform differently.

Predicting timing of expression in the *S. cerevisiae* cell-cycle using direct regulatory interactions

Previously, regulatory interactions were used to predicted gene expression in *S. cerevisiae* (Beer and Tavazoie, 2004) as well as other species (Chikina et al., 2009; Zou et al., 2011; Panchy et al., 2014; Uygun et al., 2017). Both general patterns of expression (Beer and Tavazoie, 2004) and response to specific conditions (Zou et al., 2011) have been accurately predicted, but distinguishing the phases of cycling expression patterns has proven difficult, even when the phases are associated with distinct functions and the timing of the cycle is expected to be strictly controlled (Panchy et al., 2014). Our previous attempt to identify regulators of timed expression relied primarily on computationally identified putative regulatory interactions (Panchy et al., 2014). However, we can take advantage of the nearly complete characterization of regulatory interactions in *S. cerevisiae* to address this question more directly. Because the TF-target interactions amongst our four data sets show little overlap, we cannot define a single GRN for *S. cerevisiae*. Therefore, we chose to compare the predictive power of TF-target interactions derived from different data sets and determine which are the most useful for predicting cell-cycle expressed genes.

To examine cell-cycle expression, we used cell-cycle expressed genes from Spellman et al. (1998), which are available at the Yeast Cell Cycle Analysis Project (http://genome-www.stanford.edu/cellcycle/). In this study, cell-cycle expressed genes are defined as those genes whose expression oscillates in a sinusoidal-like fashion over the cell cycle with distinct minima and maxima. These genes can be clustered into broad categories based on the timing or

"phase" of peak expression during the cell-cycle. Spellman et al. identified five such clusters of 71 to 300 genes corresponding to the G1, S, S/G2, G2/M, and M/G1 phases of the cell cycle. (**Supplementary Figure 3.2**). While it is known that each phase represents a functionally distinct period of the cell-cycle, the extent to which regulatory mechanisms are distinct or shared both within cluster and across all phase clusters is unknown. The coverage of the genes in each phase cluster by TF-target interactions from different data sets varies between 64% and 100%, but the average coverage of expression clusters is >70% for all data sets (see **Supplementary Table 3.1**), such that we expect the results of any predictor to be generalizable across the entire cluster.

In order to predict both general and phase-specific expression during the cell cycle, we used a Support Vector Machine (SVM) algorithm to classify *S. cerevisiae* genes as being cell-cycle expressed or not and, independently, classify genes as being expressed in specific phases of the cell cycle as defined in Spellman et al. (see Methods for details). The performance of each classifier was assessed using the Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC), which ranges from a value of 0.5 for a random classifier to 1.0 for a perfect classifier. To compare different types of interaction data, we used each of the five sets of TF-target interactions to independently predict expression. The AUC-ROC values for the best-performing classifiers generated by each data set are reported in **Figure 3.3**.

From the distribution of AUC-ROC values, there is an apparent relationship between performance and both the source of TF-target interactions and the timing of expression during the cell cycle. We confirmed these relationships by doing analysis of variance (ANOVA) on the performance of classifiers from each data set (see Methods). There was a significant relationship between AUC-ROC and data set ($pv < 2e-16$), expression phase ($pv < 2e-16$), and the interaction

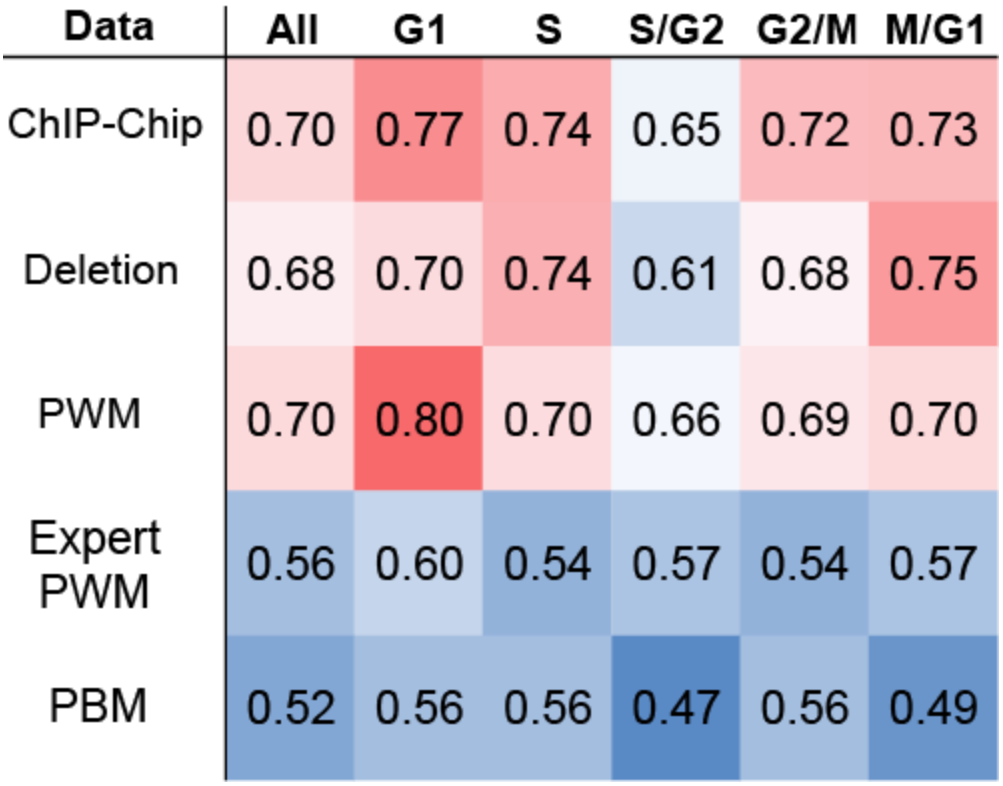


Figure 3.3 Performances of classifiers using TF-target interactions across all data sets. Heatmap of the AUC-ROC values for SVM models trained each set of cell-cycle expressed genes (all cell-cycle genes and genes expressed during the G1, S, S/G2, G2/M, or M/G1 phase) and classified using TF-target interactions derived from each feature set (ChIP-Chip, Deletion, PWM, Expect PWM, and PBM). The reported AUC-ROC for each classifier is the average AUC-ROC of 100 data sets composed of a balanced number of positive (cell-cycle genes) and negative (non-cell-cycle genes) classified using the parameters that maximize performance for that model (see Methods). Dark red shading indicates an AUC-ROC closer to 1 while dark blue indicates an AUC-ROC closer to zero.

between data and phase (pv < 2e-16). This relationship is not entirely dependent on the number of feature types, as the performance of the PWM classifier remains is unaffected if we only include features for TFs present in the ChIP-Chip data set or in Deletion data set (**Supplementary Figure 3.3A**). Similarly, if only the most important 150 features defined using SVM weights (see Methods) are included, so that the total number of features is similar to the number in the ChIP-Chip and Deletion data sets, the AUC-ROC only declines for G1 and improves slightly for S-G2 and all cyclic genes. However, if we restrict the ChIP-Chip, Deletion, and PWM to only TF features present the PBM data set (the one with the fewest TFs), we do see reduction the performance (**Supplementary Figure 3.3B**), though ChIP-Chip, Deletion and PWM still perform better than PBM, even with a reduced feature set. This indicates that, after a certain threshold, reducing the number of TFs covered by set of TF-target interactions will affect the ability to predict cyclic expression, though the magnitude of this effect is dependent on how the TF-target interactions were defined.

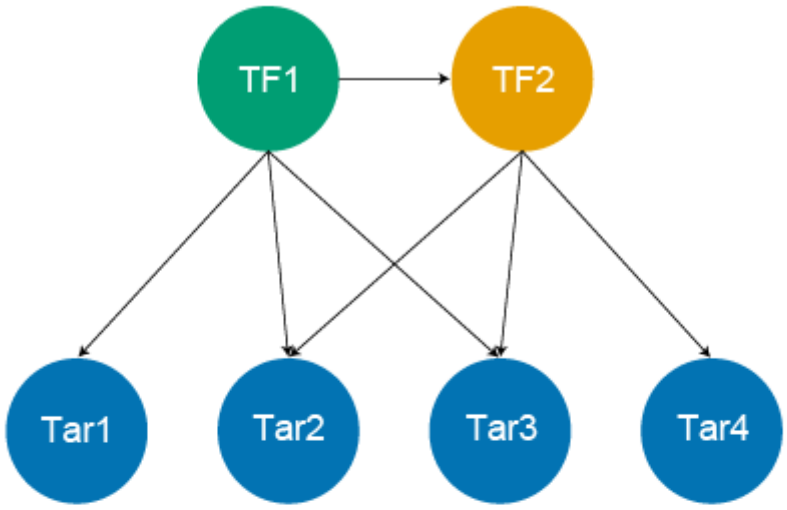
Overall, these results indicate that both cell-cycle expression in general and timing of cell-cycle expression can be predicted using direct regulator interactions, with ChIP-Chip interactions alone able to predict all clusters except S/G2 with an AUC-ROC > 0.7. While this suggests that there is significant regulatory information present in TF-target interactions that is relevant to cell-cycle expression, this information is incomplete given that our classifiers are imperfect. In particular, no set of TF-target interactions can classify S/G2 expressed genes with an AUC-ROC > 0.7. One possible explanation for this shortcoming is that this phase bridges the replicative phase (S) and the second growth phase (G2) of the cell-cycle, and therefore represents a heterogeneous set of genes with diverse functions and regulatory programs. This hypothesis is supported by the fact that S/G2 genes are not significantly over-enriched for any Gene Ontology

terms (Ashburner et al. 2000; Gene Ontology Consortium 2015) and are only significantly under-enriched for genes with mitochondrion (GO:0005739), nucleus (GO:0005643), cytoplasm (GO:0005737), and RNA binding (GO:0003723) annotations, which are also under-enriched for all expression clusters expect S-phase. Alternatively, direct regulators alone could be insufficient to characterize the regulation of genes in this phase cluster as higher-order interaction between regulators could be involved in regulation of S/G2 expression. With respect to the cell cycle and gene expression timing in general, the question is what sort of regulatory interactions would we expect to give rise to expression only at a particular time?

Predicting timing of expression during the *S. cerevisiae* cell-cycle using feed-forward loops

Given that TF-target interactions produced useful, but imperfect classifiers of cell-cycle expression, our next step was to identify interactions between TFs that can be used to improve prediction. Previously, statistical enrichment of TF-binding co-occurring amongst co-expressed genes has been used to identify regulatory interactions that are useful for prediction (Zou et al., 2011). However, there is no guarantee that these statistically significant interactions are biologically important. Instead, we decided to focus specifically on "network motifs", which are patterns of regulatory interactions that are enriched in a biological network and thus theorized to be functionally important (Alon 2007a). In particular, we chose to focus on feed-forward loops (FFLs). An FFL is a network motif that consists of a primary TF that regulates a secondary TF and a target gene that is regulated by both the primary and secondary TFs (see **Figure 3.4A**). This type of network motif is expected to result in peak expression following a delay after the expression of the primary TF is induced (Alon 2007a), and is therefore a potential regulatory mechanism for phase-specific expression in the cell-cycle. Furthermore, FFLs can be used to compose more complex interactions. For example, negative-feedback loops, which have

A



B

Data	All	G1	S	S/G2	G2/M	M/G1
ChIP-Chip	0.70	0.78	0.82	0.69	0.76	0.76
Deletion	0.68	0.74	0.70	0.68	0.71	0.75
PWM	0.66	0.72	0.69	0.66	0.69	0.65
Expert PWM	0.60	0.65	0.60	0.59	0.64	0.67
PBM	0.51	0.55	0.55	0.46	0.56	0.51

Figure 3.4 Performance of classifiers using only FFLs across all data sets (A) Representative feed-forward loops (FFLs) in a GRN. The presence of a regulatory interaction between TF1 and

Figure 3.4 (cont'd)

TF2 means that any target gene which is co-regulated by both of these TFs is part of a FFL. For example, TF1 and TF2 form a FFL with both Tar2 and Ta3, but not Tar1 or Tar4 because they are not regulated by TF2 and TF1, respectively. (B) Heatmap of AUC-ROC values for SVM classification models of each cell-cycle expression set (All cell-cycle genes and genes expressed during the G1, S, S/G2, G2,M, or M/G1 phase) using FFLs derived from each feature set (ChIP-Chip, Deletion, PWM, Expect PWM, and PBM). The reported AUC-ROC for each classifier is the average AUC-ROC of 100 data sets composed of a balanced number of positive (cell-cycle genes) and negative (non-cell-cycle genes) classified using the parameters that maximizes performance for that model (see Methods). Dark red shading indicates an AUC-ROC closer to 1 while dark blue indicates an AUC-ROC closer to zero.

previously been identified as being involved in the regulation of biological oscillations (Bertoli, Skotheim, and de Bruin 2013; Pett et al. 2016), are composed of two FFLs which identical but for the direction of the regulatory interaction between the TFs. We can potentially capture elements of more complicated regulatory pathways by identifying their constituent FFLs.

We defined FFLs in *S. cerevisiae* using the same four types of regulatory data sets used to identify TF-target interactions. In order to confirm that FFLs do represent a significantly enriched network motif in *S. cerevisiae* GRNs, we calculated the expected number of FFLs based on the total number of interactions in each GRN and the frequency of TF-TF interactions (see Methods). We compared these expected values to the actual number of FFLs in each of the five GRNs and found that in each case, more FFLs were present in the GRN than expected, indicating FFLs are, in fact, an overrepresented network motif (see **Table 3.2**). TF-TF interactions alone are highly correlated with the frequency of TFs ($r^2 = 0.93$) and the total number of TF-TF interactions ($r^2 = 0.87$) in each data set (see **Supplementary Figure 3.4**). Given that the occurrence of TF-TF interactions appears to depend on network size and TF frequency, the enrichment of FFLs indicate interacting TFs co-regulate the same target genes more often than expected by random chance.

Given that FFLs are enriched in our GRNs, we built models of cell-cycle expression using only regulation by FFLs as features. As with TF-target regulations, we treated each GRN independently because there was little overlap between data sets; 97.6% of FFLs were unique to one data set and no FFL was common to all data sets (see **Supplementary Figure 3.5**). Fewer of the 800 cell-cycle genes defined in Spellman et al. (1998) were targets of an FFL, with three of the five sets having fewer than 50% of genes covered by a FFL (see **Supplementary Table 3.2**). Hence, the models made with FFLs will likely be relevant to only a subset of cell-cycle

Table 3.2 Observed and expected number of FFLs in GRNs defined using different data sets

Data Set	Observed FFLs	Mean Expected FFLs[1]	Stdv of Expected FFLs[2]	Z-Score[3]
ChIP-Chip	3777	811	28.47	104.15
Deletion	13,162	2427	49.26	217.90
PWM	75,514	52,915	230.03	98.24
Expert PWM	1700	398	19.94	65.26
PBM	67,895	47,371	217.64	94.30

1. The mean of FFLs expected in a GRN was determined using the cube of the mean connectivity of the GRN (see Methods)
2. The standard deviation of FFLs expected in a GRN was determined using the cube of the mean connectivity of the GRN (see Methods)
3. The z-score reflects the difference between the observed and expected number of FFLs divided by the standard deviation of the expected number of FFLs (see Methods).

expressed genes, but may still be useful for identifying TF-TF interactions important for the regulation of cell-cyclic expression. Using the same machine learning approach for prediction and assessment, we found the same overall pattern of performance with FFLs as we did using direct regulators (**Figure 3.4B**). Again, the best predictions were from GRNs derived from ChIP-Chip, Deletion, and all PWMs. However, predictions using ChIP-Chip FFLs had the highest AUC-ROC values for all phases of expression. ChIP-Chip FFL models also had higher AUC-ROCs for each phase than those based on direct regulation, though it is important to note that the ChIP-Chip FFL set had a much lower coverage of cell-cycle expressed genes, 34%, compared to 82% for direct regulators. To test how restricting the set of cell-cycle genes impacts preformance, we used ChIP-Chip TF-target interactions to predict cell-cycle expression for the same 34% of cell cycle genes and found performance of predictions was improved (**Supplementary Table 3.3**) compared to using all cell-cycle genes (see **Figure 3.3**). Hence, the improved performance from FFLs may stem from the subset of cell-cycle genes being used covered by the ChIP-Chip FFL set being easier to predict using any type of regulatory feature.

Given that models based on ChIP-Chip TF-target interactions predict cell-cycle genes covered by ChIP-Chip FFLs as well as the FFL model, one might assume the information present in FFLs is redundant with TF-target interactions. However, it is important to remember that this subset of cell-cycle genes, which is easier to predict, could not be identified without using FFLs. For this reason, in spite of their limited coverage of cell-cycle genes, FFLs complement TF-target regulations, specifically by contributing the classification of the subset that they do predict well. Additionally, the results of the ANOVA analysis described in the previous section indicated that the interaction between data type and phase of expression had a significant effect on the performance of the classifier. Hence, further improvement could be gained not only by

including both direct TF-target and FFL interactions, but also by combining interaction across data sets. However, it is unlikely that all the features from any single data set are relevant to making accurate predictions, so it is necessary to distinguish between important and unimportant features before attempting to construct a classifier based on different types of TF-targets and FFL interactions from multiple data sets.

Using feature importance to merge GRNs and improve prediction of cell-cycle expression

Both the improved performance of FFLs on a subset of cyclic-genes and the effect which data set has on the performance predicting specific time suggest that a better classifier can be constructed by combining features and data sets. To do this we focused on interactions identified from the ChIP-Chip and Deletion data sets because these interactions had better predictive performance than PBM, PWM and Expert PWM interactions. Furthermore, using ChIP-Chip and Deletion GRNs are expected to be complementary because they identify interactions using independent methods: ChIP-Chip interactions represent binding in the absence of a proven change in expression, while in the Deletion data there is evidence of changes in expression, but not binding.

In order to merge regulatory information from the ChIP-Chip and Deletion GRNs, we first identified TF and TF-TF interactions that were important for each of the classifiers based on SVM weight (see Methods). Features enriched in cell-cycle expressed and non-cell-cycle expressed genes are differentiated by the sign of the weight: positive weights indicate a feature is over-enriched in cell-cycle genes while a negative weight indicates a feature is under-enriched in cell-cycle genes. Because we expect importance to vary across phases in a data set-dependent fashion, we defined the importance of each feature for each phase-specific classifier based on ChIP-Chip and Deletion data independently. We used the same criteria for importance across all

models, we selected features based on four different percentiles of SVM weight: (1) 10th percentile of positive weights, (2) 25th percentile of positive weights, (3) 10th percentile of positive and negative weights, (4) 25th percentile of positive and negative weights (see Methods). Using this approach allowed us to assess if accurate predictions only require cell-cycle associated (i.e. positive weight) features, or if performance depends on exclusionary (i.e. negative weight) features as well.

Before combining features selected using the above criteria, we first assessed the predictive power of each subset of TF-target or FFL features, separately and combined, for both ChIP-Chip (**Figure 3.5A**) and Deletion (**Figure 3.5B**) interactions. We found the same overall pattern of performance as previous classifiers; classifiers built using ChIP-Chip FFL subsets outperformed classifiers from ChIP-Chip direct interactions across all phases, while the performance of classifiers using Deletion TF-target interactions and FFLs varied depending on the phase, with FFL classifiers performing better with S/G2 and G2/M genes like before. For all subsets consisting either entirely of TF-target regulations or FFLs, the 25th percentile of both positive and negative SVM weights performed best, except for Deletion FFL predictions for the S/G2 phase. While this would seem to suggest that more features leads to better performance, these 25th percentile subsets perform equally well or better than the full data set for both TF-target interactions and FFLs with a few exceptions (Deletion direct interactions for G2/M and both ChIP direct interactions and FFLs for G1) (see **Figures 3.3** and **3.4B**). Similarly, when combining direct regulators and FFLs, the 10th percentile of positive and negative SVM weights had the best performance in 75% of cases. These results indicate that we can achieve equal or improved performance predicting cell-cycle expression using a subset of important features, so long as both features associated with cell-cycle and non-cell-cycle gene expression are included.

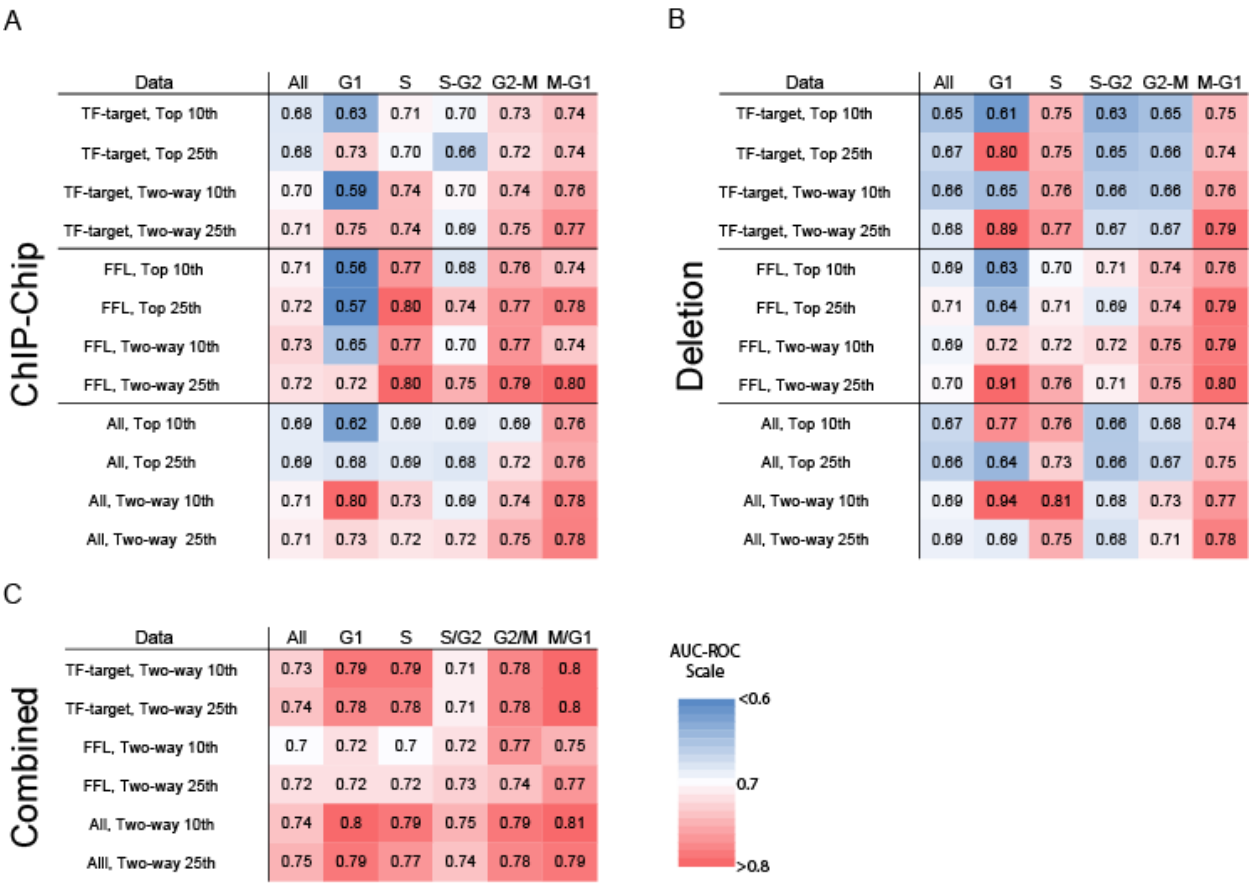


Figure 3.5 Performance of classifiers built using important features from ChIP-Chip, Deletion, and combined ChIP-Chip/Deletion data set. (A) Heatmap of AUC-ROC values for SVM classification models for each cell-cycle expression set (All cell-cycle genes and genes expressed during the G1, S, S/G2, G2, M, and M/G1 phases) constructed using a subset of ChIP-Chip TF-target interactions, FFLs, or both. Subsets of features were defined using the importance of features (either TFs or TF-TF interactions) as follows: features in the top 10th percentile of importance (Top 10th), in the top 25th percentile of importance (Top 25th), the top and bottom 10th percentiles of importance (Two-way 10th), and the top and bottom 25th percentiles of importance (Two-way 25th) (see Methods). The reported AUC-ROC for each classifier is the average AUC-ROC of 100 data sets composed of a balanced number of positive (cell-cycle genes) and negative (non-cell-cycle genes) examples classified using parameters that maximize

Figure 3.5 (cont'd)

performance for that model (see Methods). Because of the wide range of values, the heatmap is scaled such that the darkest blue value indicates an AUC-ROC <0.6 and the darkest red color indicates an AUC-ROC > 0.8. (B) Heatmap of AUC-ROC values for SVM classification models for each cell-cycle expression set (All cell-cycle genes and genes expressed during the G1, S, S/G2, G2, M, and M/G1 phases) constructed using a subset of Deletion TF-target interactions, FFLs, or both. Subsets of genes are defined as in (A). AUC-ROC was calculated and the heatmap colored as in (A). (C) Heatmap of AUC-ROC values for SVM classification models of each cell-cycle expression set (All cell-cycle genes and genes expressed during the G1, S, S/G2, G2, M, and M/G1 phases) constructed using a subset of TF-target interactions, FFLs, or both from combined ChIP-Chip and Deletion data. Subsets of features are defined as in (A) expect that only the Two-way 10^{th} and Two-way 25^{th} cutoffs are used and they are applied to both the ChIP-Chip and Deletion data sets. AUC-ROC was calculated and the heatmap colored as in (A).

The final models were built by combining ChIP-Chip and Deletion features, including subsets with both positive and negative weights. The total number of features in each subset can be found in **Table S4**. Although some cell-cycle genes are regulated by a TF-target interaction but not an FFL, to ensure that the results are comparable, we used all cell-cycles genes covered by at least one TF-target interaction from either data set to assess the combined ChIP-Chip/Deletion models. As such, it is not unexpected that the performance of combined models using FFLs was lower compared with those using TF-target interactions (**Figure 3.5C**). Nevertheless, the AUC-ROC of the combined FFL models were > 0.70 for all phase clusters (**Figure 3.5C**) and, except for G1, outperformed predictors based on any full set of TF-target interactions (**Figure 3.3**). Furthermore, the combined FFL models outperformed combined TF-target interaction models in predicting cell-cycle gene expression during S/G2. Both 25th percentile models had similar precision (96.0%), but the FFL model had higher recall of cell-cycle expressed genes (49.3%) than the TF-target interaction model (44.9%). These two models also correctly identified slightly different subsets of S/G2 expressed genes, with seven correctly predicted only by the FFL model and four correctly predicted only by the TF-target interaction model. Hence, it is not surprising that using both TF-target interactions and FFLs showed the best performance for all phases of cell cycle expression (**Figure 3.5**).

Overall, the consistency with which classifiers built using both ChIP-Chip and Deletion data outperform classifiers built with just one data type indicates the power of using complementary characterizations of a GRN to predict expression. Furthermore, these combined models outperform classifiers based on single data sets even though they contain fewer total features. The performance of features which were found to be important in one of our original models, both alone and in combination, not only indicates that feature selection can be a

powerful tool for improving gene expression predictions, but also that features with high importance may be enriched in TF and TF-TF interactions that are specific to the control of cell-cycle expression. While we would expect that many TFs with high importance are cell-cycle regulators, we may also use important TF-target and TF-TF interactions to discover novel TF functions that are associated with cell-cycle regulation.

Functions of TFs important for predicting cell-cycle expression

In our analysis of the ChIP-Chip and Deletion data sets, we found that performance of classifiers could be maintained while including only the most important features. To test if the selecting for features important to predicting cell cycle expression identifies true biological regulations, we asked whether the 10th and 25th percentile of TFs features were enriched in cell cycle-related genes. Of the 25 TFs that have been annotated as cell-cycle regulators in *S. cerevisiae* (GO:0051726), 20 were identified as features important to predicting cell-cycle expression in either the ChIP-Chip or Deletion data set. For ChIP-Chip, the 10th percentile of the most important TFs from all phases except M/G1 is enriched for cell-cycle genes, while for the 25th percentile of important TFs, only the features of general classifier (i.e. cell-cycle genes from all phases) are enriched for cell-cycle genes (Fisher's Exact Test, **Supplemental Table 3.5**). The pattern of enrichment was less clear for the Deletion interactions. While important Deletion TFs from either the 10^{th} or 25^{th} percentile are not enriched for cell-cycle genes, the top three most important TFs from the general classifier are annotated as cell-cycle genes (**Supplemental Table 3.5**). Furthermore, the majority of the 25 cell-cycle annotated TFs are present in the 25th percentile of important features in at least one phase of cell cycle in both the ChIP-Chip (14) and Deletion (13) data sets. To summarize these findings, the important features from our classifiers tend to be associated with the cell-cycle, which suggests our data accurately

represents the true *S. cerevisiae* GRN and that our predictive methodology correctly identified associations between regulators and expression patterns.

While the set of important TFs identified by our classifiers are enriched for cell-cycle TFs, these TFs still represent the minority of important TFs. To better understand the functions of these other important TFs, we looked for additional enriched GO Terms in the 10th and 25th percentile of important TFs from both the ChIP-Chip and Deletion data sets (**Supplemental Table 3.6**). We found 124 GO terms over-represented and 5 under-represented in at least one feature set, with no terms being over-represented in one set and under-represented in another. One GO term, mitochondrion (GO:0005739), was under-represented in all four feature sets while cytoplasm (GO:0005737) and membrane components (GO:0016020 and GO:0016021) were under-represented in both 25th-percentile feature sets. This was expected given that all features are TFs. There were 19 GO terms over-represented in all four features sets, including several generic TF functions (e.g. transcription, DNA-templated, regulation of transcription, DNA-templated , DNA binding), but also more specific functions including the positive regulation of transcription in response to variety of stress conditions (e.g. salt, starvation, freezing; **Supplemental Table 3.6**). This association is not without precedent, as a previous study found that cell-cycle genes, particularly those involved in the G1-S phase transition, are needed for heat-shock response (Jarolim et al. 2013). However, our results indicate a much broader overlap between cell-cycle regulation and stress response.

The majority of over-enriched GO terms were unique either to ChIP-Chip features (45) or Deletion features (29). In general, ChIP-Chip TFs were over-represented for terms related to regulation of growth and phenotype switching while Deletion TFs were over-represented for terms related to metabolism and the regulation of ribosomes. The full list of GO terms and the

features sets they are enriched in can be found in **Supplemental Table 3.6**. Though there is a large degree of overlap in the TFs present in the ChIP-Chip and Deletion sets, the difference in TF-target interactions results in different subsets of these TFs being identified as the most important, and therefore differential enrichment of gene function amongst the best features of each set. In particular, the best features derived from ChIP-Chip interactions were over-enriched for growth related functions, while the features derived from Deletion were enriched for regulation of metabolism. The distinct functions of important TFs from the ChIP-Chip and Deletion data supports the hypothesis that the improvement in predictive power from combining feature sets was due to the distinct, but complementary characterization of gene regulation in *S. cerevisiae*.

Finally, we identified GO annotations enriched in TFs important for predicting individual phases of cell-cycle expression. Because we previously identified GO terms enriched in all regulators of the cell-cycle, we specifically looked for terms that were not only robust to data set and importance threshold, but also unique to a single phase of the cell cycle. Out of 274 terms enriched in at least one phase of the cell cycle, 94 were unique to a single phase (see **Supplemental Table 3.7)**, but only one was enriched in all four data sets (selenite ion response, GO:007271, in G2M). An additional 20 unique GO terms were enriched in all ChIP-Chip feature sets and 4 were in all Deletion feature sets; however there were 60 GO terms whose enrichment in more than one phase is supported by multiple feature sets. This indicates that the regulation of expression timing across the cell-cycle involves a certain degree of overlap and that we should be able to find examples of TFs that are important for multiple phases of cell-cycle expression. We theorize that such "general" regulators of cell-cycle expression are central to GRNs specific

to cell-cycle regulation. To test this hypothesis we made use of the importance of both individual TF-target and regulatory TF-TF interactions to characterize the structure of cell-cycle regulation.

Identifying regulatory modules for cell-cycle expression

Although looking at the importance and functional enrichment of individual TFs is important for identifying the factors and processes important to the timing of cell-cycle expression, ultimately what we want understand is how regulatory interactions play a role in determining time-specific expression across the cell-cycle. In particular, the prominence of overlapping enriched functions across the cell-cycle suggests that there may be groups or "modules" of TFs responsible for regulating multiple phases of expression. This is supported by the observation that 7.9 and 10.6% of TFs are important for >1 phase at the 10th percentile cutoff and 32.2% and 30.4% are important for >1 phase at the 25th percentile cutoff for ChIP-Chip and Deletion interactions, respectively. However, when we hierarchically clustered TF interactions based on their importance for general and cell-cycle phase specific classifiers (**Supplemental Figure 3.6**) we found no large clusters of TFs that could be responsible for regulating expression at multiple phases. One possible explanation for this could be that, although both positive and negative importance features were necessary to construct a good predictor of cyclic expression, using the full range of importance values for all features is confounding. For example, if a module was important for regulating expression at M/G1 and G1, we would expect that the importance scores for TFs in that module would be highly correlated during those phases, but could vary from slightly positive to very negative in the other phases. Thus, different criteria are required for defining potential regulatory modules.

In order to identify regulatory modules without relying on correlated importance scores, we used TF-TF interactions to build a network of regulators. To begin, we filtered the set of

ChIP-Chip TF-TF interactions by the 10th percentile of importance for predicting cell cycle genes (**Figure 3.6A**). We then identified TF-TF interactions that were above the 10th percentile of importance for one or more phases and found that 61% of all interactions important for predicting cell-cycle genes were also important for predicting at least one phase cluster and 34.8% were within the top 10th percentile for >1 phases. All but one of the interactions important for predicting phase-specific expression were concentrated around four groups of genes (colored regions, **Figure 3.6A**). Two of these groups, Swi6-Swi4-Mbp1 (red), which is a regulator of the G1/S phase transition, (Iyer et al. 2001; Wittenberg and Reed 2005; Bean, Siggia, and Cross 2005) and Fkh1-Fkh2-Ndd1, which is involved in the regulation of S/G2 (G. Zhu et al. 2000) and G2/M (Koranda et al. 2000) expressed genes, are known regulatory complexes. Therefore, it is not surprising that interactions amongst these groups are primarily important for early (G1 through S/G2) and middle (S to G2/M) phases of cell-cycle expression, respectively. In summary, we were able to identify regulatory modules important for predicting multiple expression phases that are made up of regulatory complexes known to be important for cell-cycle progression.

We also found interactions important for multiple phases of cyclic-expression that are not part of canonical cell-cycle regulatory complexes. For example, the feedback loop between Ste12 and Tec1 was identified in our models as an important regulator of gene expression during S/G2 and M/G1. (purple, **Figure 3.6A**) Ste12 and Tec1 are known form a complex that shares co-regulators with Swi4 and Mbp1 to promote filamentous growth (van der Felden et al. 2014), one of the functions enriched amongst TFs important for predicting cell cycle expression. However, neither of these TFs interacts directly with the Swi6-Swi4-Mbp1 complex (van der Felden et al. 2014) nor are they part of the annotated set of cell-cycle regulators. Similarly, interactions

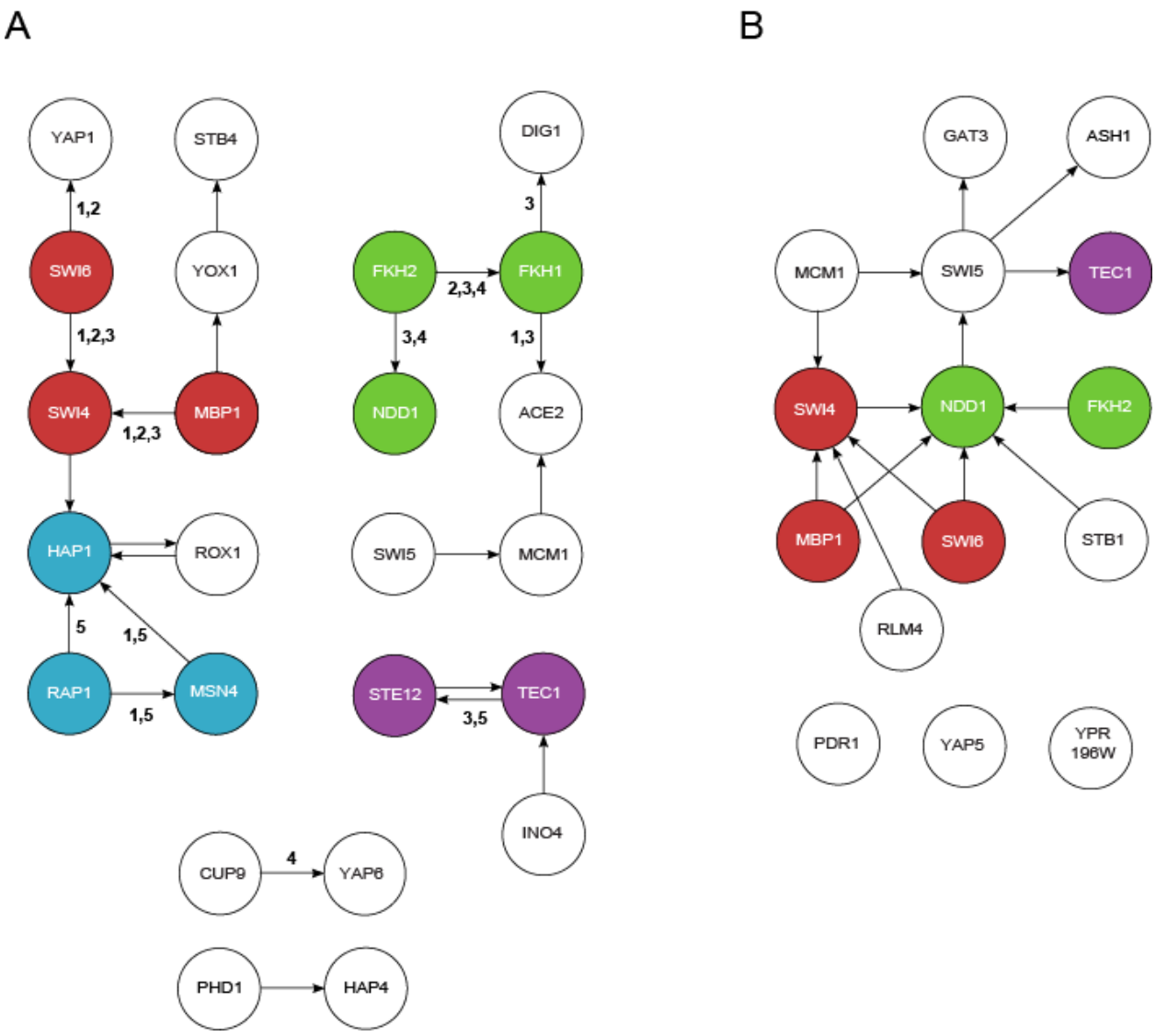


Figure 3.6 The cell-cycle expression GRN defined using the 10th percentile of ChIP-Chip features. (A) A network of ChIP-Chip TF-TF interactions selected from the ChIP-Chip GRN constructed using the ChIP-Chip FFLs from the top 10th percentile (see Methods) of importance for predicting all cell-cycle expressed genes. Interactions are further annotated with the stage of cell-cycle expression (1 = G1, 2 = S, 3 = S/G2, 4 = G2/M, 5 = M/G1) they are important for predicting (10th percentile of SVM weight in ChIP-Chip models). Four modules with interactions important for predicting >1 phase of cell-cycle expression are highlighted by color: Swi6-Swi4-Mbp1 (red), Fkh2-Fkh1-Ndd1 (green), Ste12 and Tec1 (purple) and Rap1-Msn4-Hap1 (blue).

Figure 3.6 (cont'd)

(B) A network of TF-TF interactions from the ChIP-Chip GRN which exists amongst TFs in the top 10th percentile of importance for predicting all cell-cycle expressed genes using ChIP-Chip TF-target interactions. Genes are colored as in (A).

between Rap1, Hap1, and Msn4 are in the top 10th percentile of important ChIP-Chip TF-TF features for predicting the M/G1 and G1 phases (blue, **Figure 3.6A**). However, none of these TFs are annotated cell-cycle regulators; Rap1 is involved in telomere organization (Guidi et al. 2015; Laporte et al. 2016), Hap1 is an oxygen response regulator (Keng 1992; Ter Linde and Steensma 2002), and Msn4 is a general stress response regulator (8641288, 8650168).

Of the genes involved in complexes that regulate multiple phases of the cell cycle, but that are not annotated cell-cycle regulators, only Tec1 is found in the 10th percentile of important TF features in ChIP-Chip data. Furthermore, Rap1 and Hap1 are not found in any important TF feature set. Rather their SVM weights are near zero, suggesting that direct regulation by Rap1 or Hap1 is not significantly associated with either cell-cycle expression or non-cell-cycle expression. Hence, it is only by looking at interaction between regulators that the importance of these TFs becomes apparent. In addition, had we only considered interaction amongst the 10th percentile of TF features with the best performance in ChIP-Chip data (**Figure 3.6B**), we would missed the Ste12-Tec1 and Rap1-Hap1-Msn4 modules entirely. Additionally, Fkh1 is not amongst the 10th percentile of TF features, so part of the Fkh1-Fkh2-Ndd1 module would have been overlooked as well. The network of all interactions amongst the 10th percentile of TFs also includes many TF-TF interactions involving Ndd1 and Swi5 that were not found to be important to predicting cell-cycle expression in our classifiers. The source of TF-interactions is significant as performing the same analysis on the 10th percentile of TF-TF interactions in the Deletion data set revealed none of the same modules as in the ChIP-Chip networks (**Supplemental Figure 3.7**). This includes all of the canonical interactions between Fkh1-Fkh2-Ndd1 as well as the interaction between Swi6-Mbp1. This illustrates the power of identifying potential TF-TF

interactions in a way that is independent of individual TF importance, though the results are highly dependent on how such interactions are defined.

To further investigate important regulatory interactions, we expanded our network to include the 25th percentile of TF-TF interactions from ChIP-Chip data (**Figure 3.7**). In the resulting network, 38 of 46 TFs (82.6%) formed a single network, while only 0.8% of networks formed by randomly drawing the same number of interactions from ChIP-Chip data had a similar or greater degree of interactivity. Comparably, 57 of 67 TFs (85.1%) of the 25th percentile of TF-TF interactions in the Deletion data set are interconnected, but 28.6% of random networks of equal size have a similar or greater degree of connectivity. We again identified the interactions of the expanded ChIP-Chip network using the 25th percentile of importance across expression phases, which resulted in 89% of the TF-TF interactions being significant in at least one phase, an increase from 61% in the 10th percentile network, but the frequency of interactions important for >1 phase remained about the same (35.6%). We should also note that the interaction between Swi4 and Mcm1, which fell just below the 25th percentile cutoff of importance for predicting general cell cycle expression, was above the 25th percentile cutoff for all phases except for G1, making it the only near universal regulatory interaction observed in this study. The majority of the new interactions important for >1 phase originated from one of the four multi-phase modules identified in the network built from 10th percentile interactions, while the remainder are distributed through the rest of the network (Cup9-Yap6, Ino4-Met4, and Met32-Ume6). Therefore, the modular structure identified in the previous ChIP-Chip network appears to be robust to the threshold we used to define importance.

Overall, the structure of the GRN built from the ChIP-Chip network indicates the presence of multiple, broad regulatory modules that interact with each other and with peripheral,

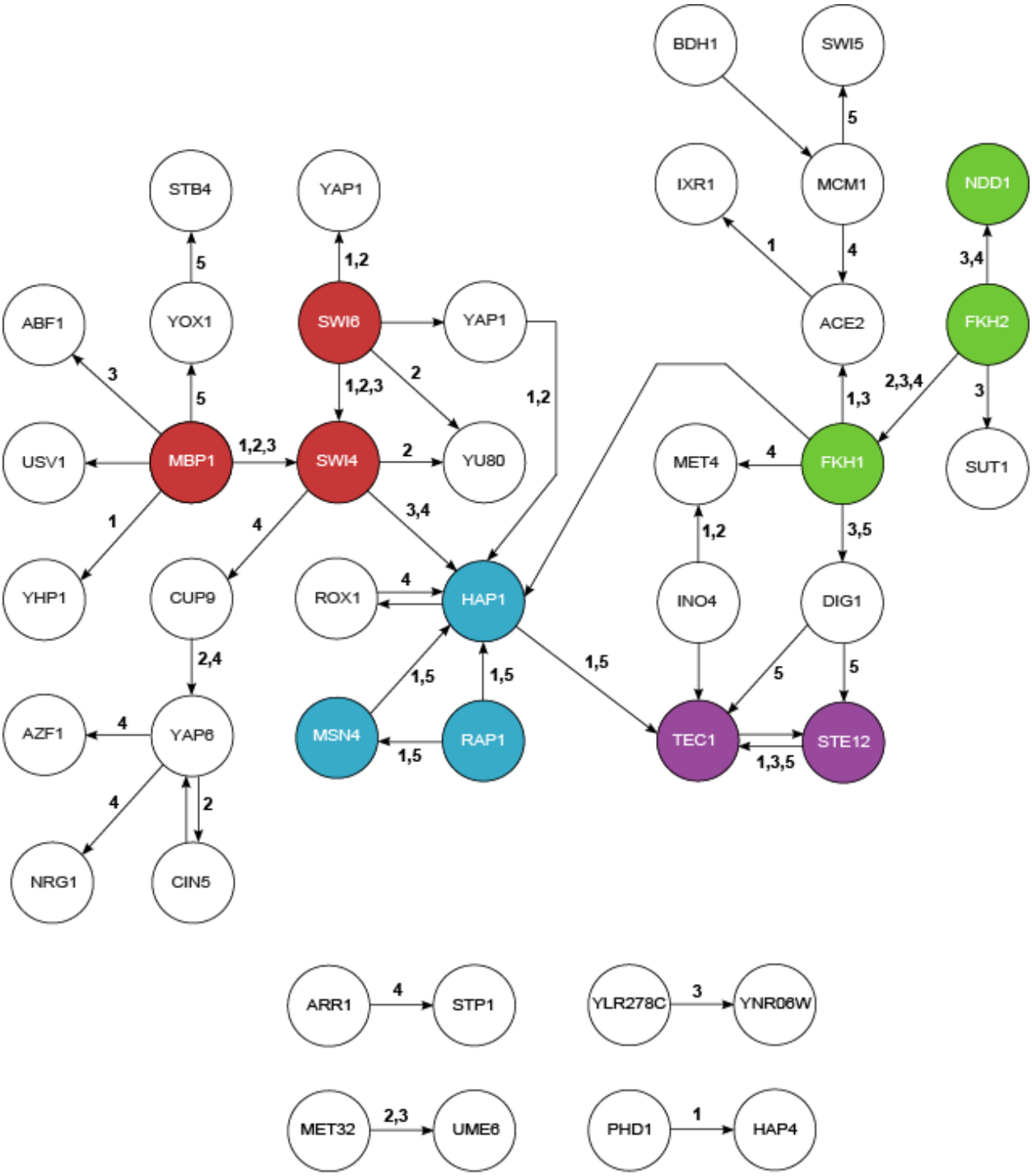


Figure 3.7 The cell-cycle expression GRN defined using the 25th percentile of ChIP-Chip TF-TF interactions. A network of TF-TF interactions selected from the ChIP-Chip GRNs constructed using ChIP-Chip FFLs from the top 25th percentile (see Methods) of importance for predicting all cell-cycle expressed genes. Interactions are further annotated with the stage of cell-

Figure 3.7 (cont'd)

cycle expression (1 = G1, 2 = S, 3 = S/G2, 4 = G2/M, 5 = M/G1) they are important for predicting (25^{h} percentile of SVM weight in ChIP-Chip models). Four modules with interactions important for predicting >1 phase of cell-cycle expression are highlighted by color: Swi6-Swi4-Mbp1 (red), Fkh2-Fkh1-Ndd1 (green), Ste12 and Tec1 (purple) and Rap1-Msn4-Hap1 (blue).

phase-specific regulators to control expression timing across the cell-cycle. Importantly, this is only true of the network built from TF-TF interactions in the ChIP-Chip feature set, while the network derived from Deletion TF-TF interactions lacks the same modularity). Differences in network structure are not unexpected, given that interactions derived from the ChIP-Chip data are inferred using direct binding to target promoters, while those from Deletion data include any target whose expression is affected by the loss of the TF, whether it binds directly or acts indirectly through another gene. Hence, we interpret the contrasting results from these data sets to mean that the direct regulation of cell-cycle expression timing involves the regulatory modules identified in the ChIP-Chip network, while there is another set of regulators identified in the Deletion network whose net effect on transcription through both direct and indirect regulatory interactions is also important for timing of expression during specific phases.

CONCLUSIONS

Predicting the expression of genes from their regulatory elements remains a challenging exercise, but one that can be useful for studying how organisms respond to various stimuli and how that response is regulated at the molecular level. Here, we have shown that the problem of predicting complex expression patterns, such as the timing of expression across the cell-cycle, is tractable using a variety of experimental and computational methods of defining TF-target interactions. In spite of painting distinctly different pictures of the *S. cerevisiae* GRN, interactions inferred from ChIP-Chip, Deletion and PWM data sets were useful for predicting genes expressed during the cell cycle and for distinguishing between genes expressed at different phases. In fact, because some cell-cycle genes were only correctly predicted using ChIP-Chip or Deletion data, integrating interactions from both data sets into a single model improved the overall accuracy of machine learning models. Furthermore, we found that models were improved with the addition of TF-TF interactions in the form of FFLs and that a subset of the most important interactions, combined with a subset of the most important TF-target interactions, performed better than either the full set of TF-target interactions or FFLs.

By studying the TFs involved in the most important TF-target interactions and FFLs we were able to infer that these interactions play a biologically significant role in regulating the cell-cycle. Using GO analysis, we found that the 10th percentile of important TFs from every phase except M/G1 were enriched for TFs with cell-cycle annotations. For the M/G1 phase we identified important TF-TF interactions that involve non-canonical cell-cycle regulators, such as the regulatory modules Ste12-Tec1 and Rap1-Msn4-Hap1. The Rap1-Msn4-Hap1 module stands out in that, while these regulators are individually poor predictors of cell-cycle expressions, interactions between these TFs are among the best predictors of both cell-cycle expression in

general and of the M/G1 and G1 phases in particular. Our GO analysis also indicated that TFs important for predicting cell-cycle expression were enriched for genes associated with metabolism, invasive growth, and stress responses, which was reflected in the network analysis as we found that interactions important for >1 phases of cell-cycle expression were clustered around TFs involved in those processes (Cst6, Ste12-Tect, Rpn4, Rap1-Msn4-Hap1).

Even though our best performing data has nearly complete coverage of the *S. cerevisiae* transcriptome, our models do not provide a complete picture of the regulation of cell-cycle expression. In particular, kinases and the interaction between kinases and TFs are known to play a key role in regulating the timing of the cell cycle, and FFLs are frequently observed in this TF-kinase network (Csikász-Nagy et al. 2009). Better characterization of TF binding sites will also help provide more accurate representation of the GRN regulating expression timing, such as novel methods of characterizing binding sites that incorporate information about both position and DNA modification (Csikász-Nagy et al. 2009; O'Malley et al. 2016). Nevertheless, this work shows that predictive models can provide a framework for identifying both regulators and regulatory interactions with biological significance to processes of interest. Understanding the molecular basis of the timing of expression is of interest not only to the cell-cycle, but other important biological processes, such as response to environmental cues, including acute stresses like predation and infection as well as cyclical changes in the environment such as light and heat. Furthermore, the approach described here is not limited to the study of expression timing, but can also be applied to any expression pattern with discrete phases.

MATERIALS AND METHODS

TF-target interaction data and regulatory cite mapping

Data used to infer TF-target interactions in *S. cerevisiae* were obtained from the following sources: ChIP-Chip (Harbison et al. 2004) and Deletion (Reimand et al. 2010) data were downloaded from ScerTF (http://stormo.wustl.edu/ScerTF/), PWMs (de Boer and Hughes 2012) and the expert curated subset of these PWMs were downloaded from YetFaSCO (http://yetfasco.ccbr.utoronto.ca/), and PBM binding scores were taken from Zhu et al. (see Supplemental Table 5, (C. Zhu et al. 2009). For ChIP-Chip and Deletion data, the interaction between TF and their target genes were directly annotated, however, for PWMs and PBMs data we mapped inferred binding sites to the promoters of genes in *S. cerevisiae* downloaded from Yeastract (http://www.yeastract.com/). All position weight matrices were mapped for the PWM data set, however for PBM data we only used the oligonucleotides in the top 10th percentile of scores for every TF. This threshold was determined using a pilot study which found that using the 10th percentile as a cutoff maximized performance of prediction using PBM data. Mapping was done according to the pipeline previously described in Zou et al. (2011) using a threshold mapping p-value of 1e-5 to infer a TF-target interaction.

Overlap between TF-target interaction data

To evaluate the significance of the overlap in TF-target interactions between different GRNs, we compared the observed number of overlaps to what we expected were the genes regulated by each transcription factor randomized. In detail, for each set of TF-target interactions we replaced the target gene of each interaction with one that was randomly drawn from the total set of target genes across all data sets, such that the number of interactions for each TF were

preserved. For each randomization of target gene, the number of overlapping features between each pair of data set was calculated. This process was repeated 1000 time to determine the mean and standard deviation of overlap between each data set expected under this randomization regimen. To determine to degree to which our observation differed from the expectation under this random model, we applied the two-tailed z-test to the differences between the observed number of overlaps and the distribution of overlaps from the randomized trials.

Expected feed-forward loops in *S. cerevisiae* regulatory networks

FFLs were defined in each set of TF-target interactions as any pair of TFs with a common target genes where a TF-target interaction also existed between one TF (the primary TF) and the other (the secondary TF) which, for clarity, we refer to as a TF-TF interaction. The expected number of FFLs in each data set was determined according to the method described by Uri Alon in "An Introduction to Systems Biology" (Chapter 4, 2007b). Briefly, the expected number of FFLs (N_{FFL}) in a randomly arranged GRN is approximated by the cube of the mean connectivity (λ) of the network with a standard deviation equal to the square-root of the mean. Therefore, for each data set we compared the observed number of FFLs to the expected number of FFLs from a network with the same number of connections, but with those connections randomly arranged by defining λ as the number of TF-target interactions divided by the total number of nodes (TFs + target genes) and calculating mean the standard deviation as above.

Validating FFLs in cell-cycle expression

FFLs were validated in the context of cell-cycle expression by modeling the regulation and expression of genes involved in the FFL using a system of ordinary differential equations:

$$\Delta\begin{pmatrix} S \\ T \end{pmatrix} = \begin{pmatrix} \alpha_S & 0 \\ \beta_{S,T} & \alpha_T \end{pmatrix}\begin{pmatrix} S \\ T \end{pmatrix} + \begin{pmatrix} \beta_{P,S} \\ \beta_{P,T} \end{pmatrix} f(t)$$

Where S and T are the expression of the secondary TF and target gene respectively, $\propto_S$ and $\propto_T$ are the decay rates of the secondary TF and target gene respectively, and $\beta_{S,T}$ indicates the production rate of the target gene dependent on the secondary TF. In the nonhomogeneous term portion of the equation, $\beta_{P,S}$ and $\beta_{P,T}$ are the production rate of the secondary TF and target gene, respectively, which depend on the primary TF, while f(t) is the expression of the primary TF over time which is independent of both the secondary TF and the target gene. This system was solved in Maxima (http://maxima.sourceforge.net/index.html). For each FFL, maximum likelihood estimation, implemented using the bbmle package in R (https://cran.r-project.org/web/packages/bbmle/index.html), was used to fit the model parameters to the observed expression of genes during the cell-cycle as defined by Spellman et al. (1998). Each run was initialized using the same set of initial conditions and only FFLs for which a reasonable ($\propto < 0$, $\beta s > 0$), non-initial parameters could be fit were kept. Between 80 and 90% of FFLs in each data set passed this threshold, while only 21% of FFLs built from random TF-TF-target triplets were fit.

Classifying cell-cycle genes using machine learning

Predicting cell-cycle expression and phase of cell-cycle expression was done using the Support Vector Machine (SVM) algorithm implemented in Weka (Hall et al. 2009). For each SVM run, the full set of positive instances (either cell-cycle genes or genes expressed at a certain phase of the cell-cycle) and negative instances (genes in the Spellmen et al. expression data set which were not cell-cycle expressed) was used to generate 100 balanced (i.e. 1-to-1 ratio of positive to negative) inputs. Genes were only selected for the input of a SVM run if at least one interaction feature was involving that gene was present. Features consist of the presence of

regulation by a TFs FFL, or a combination of both from one or more regulatory data sets (ChIP-Chip, Deletion, PWM, Expert-PWM, PBM).

Each balanced input set was further divided into 10-folds for cross validation. SVM uses the training data to define a linear classifier (i.e. a hyperplane) in the space defined by features, which is then used to classify positive and negative instances in the test set. Each run was optimized using a grid search of two parameters: the minimum distance between the positive and negative groups (C) and the ratio of negative to positive examples in the training set (R).The tested range of each parameter was as follows: C = (0.01, 0.1, 0.5, 1, 1.5, 2.0) and R = (0.25, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4). For each pair of parameters, performance was measured using the AUC-ROC values averaged across the 100 balanced input sets. For each choice of positive class and feature set, the pair of grid search parameters which maximized the average AUC-ROC was used to define the representative model for that predictor and calculate the reported AUC-ROC for that predictor.

Evaluating the relationship between model performance, class and feature

The effect of the phase (general cell-cycle, G1, S, S/G2, G2/M or M/G1) of expression being predicted (class) and the data set (ChIP-Chip, Deletion, PWM, Expert PWM or PBM) from which TF-target interactions were derived (feature) on the performance of each SVM model was evaluated using analysis of variance (ANOVA). This was done using the "aov" function in the R statistical language using the following model:

$$S = C + D + C * D$$

Where "S" is the representative AUC-ROC score of the SVM model, "C" is a categorical feature representing the positive-class set (cyclic expression or a specific phase of expression), and "D" is a categorical feature representing the data set of regulations used.

Importance of features to predicting cell-cycle expression

The importance of a feature for each model was determined by rerunning each SVM model using the best pair of parameters with the options "-i -k" in order to generate an output files with class and features statistics. From the resulting output file, custom Python scripts were used to extract the weight value for each of the features used in the linear classifier. Features were then ordered by their weight to determine importance, such that the feature with the largest positive value (most strongly associated with the positive class) had the highest rank and the feature with the largest negative value (most strongly associated with the negative class) had the lowest rank. Because multiple features often had the same weight value, we defined cutoff scores for the 10th and 25th percentile conservatively, such that the cutoff for the Xth percentile of positive features was smallest weight above which includes X% or less of all features and the Xth percentile of negative features was the largest weight below which includes X% or less of all features. The effect of this is observed most prominently in the 25th percentile features sets as ties between feature weights were more common towards the middle of the weight distributions.

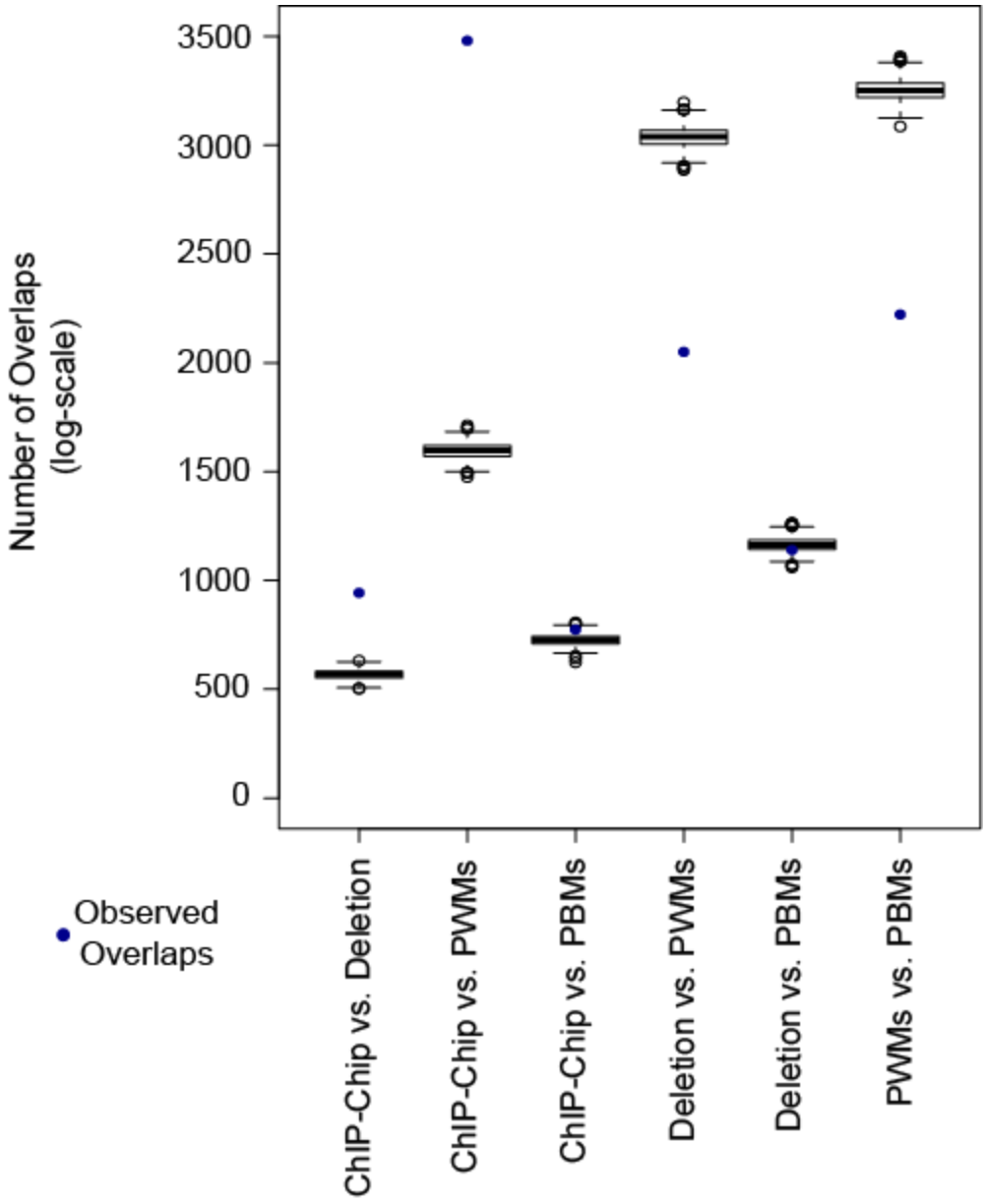
GO Analysis

GO annotation for genes in *S. cerevisiae* were obtained from the Saccharomyces Genome Database (http://www.yeastgenome.org/download-data/curation). The significance of enrichment of a particular term in a set of important TF was determined using the Fisher's Exact Test and adjusted for multiple-hypothesis testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

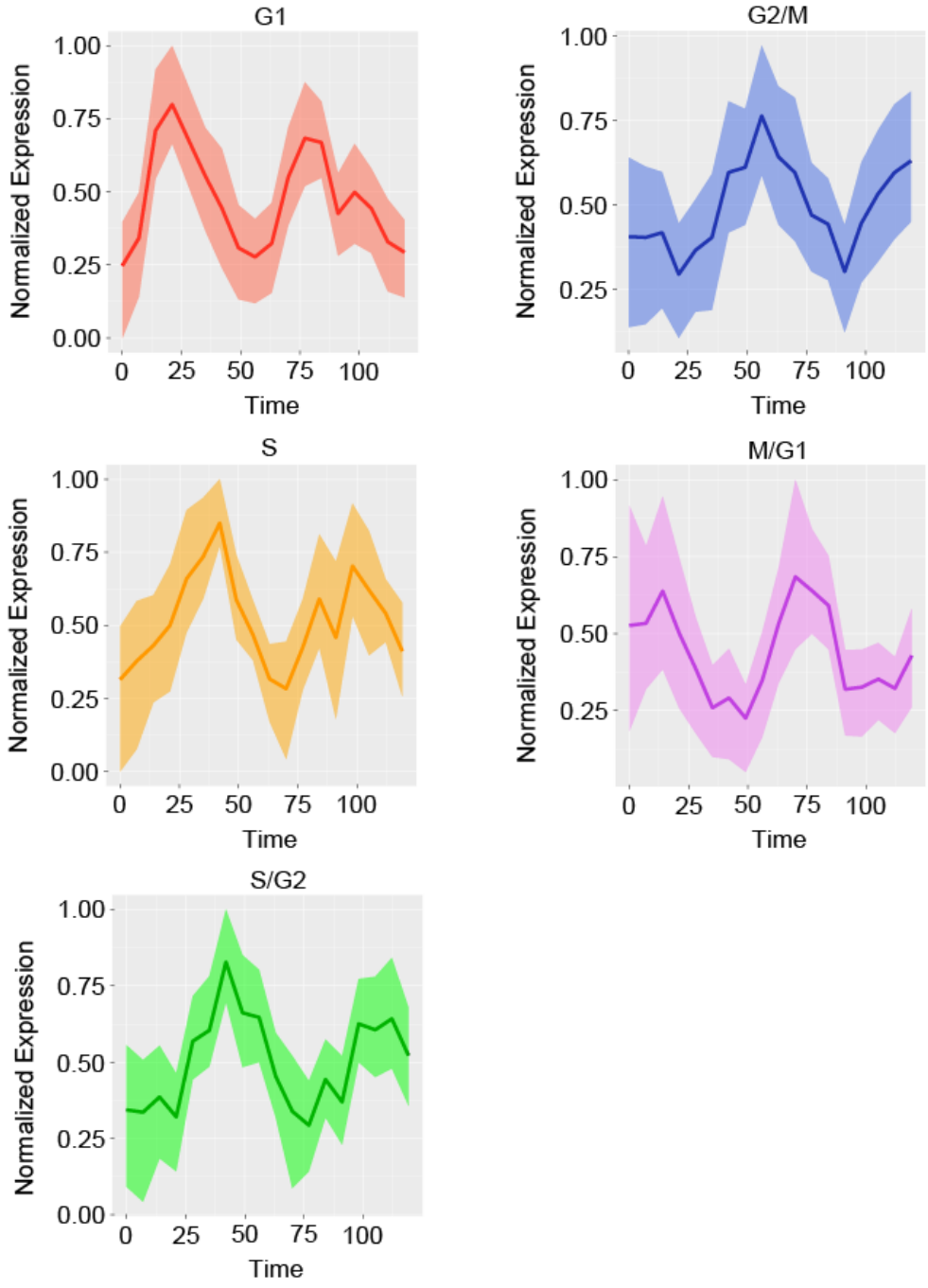
ACKNOWLEDGEMENTS

We thank Christina Azodi and Melissa Lehti-Shiu for their assistance editing this manuscript.

APPENDIX



Supplemental Figure 3.1 Expected overlaps of TF-target interactions across regulatory data sets. IQR plots of the expected number of overlapping TF-target interactions between each pair of GRNs based on randomly drawing TF-target interactions from the total pool of interactions across all data sets (see **Methods**). Blue points indicate the observed number of overlaps between each pair of GRNs.



Supplemental Figure 3.2 Expression profiles of genes expressed at specific phases of the cell-cycle. Expression profiles of genes expressed at each phase of the cell-cycle: G1 (red), S (yellow), S/G2 (green), G2/M (blue), and M/G1 (purple). Time (x-axis) is expressed in minutes and, for the purpose of display, the expression (y-axis) of each gene was normalized between 0

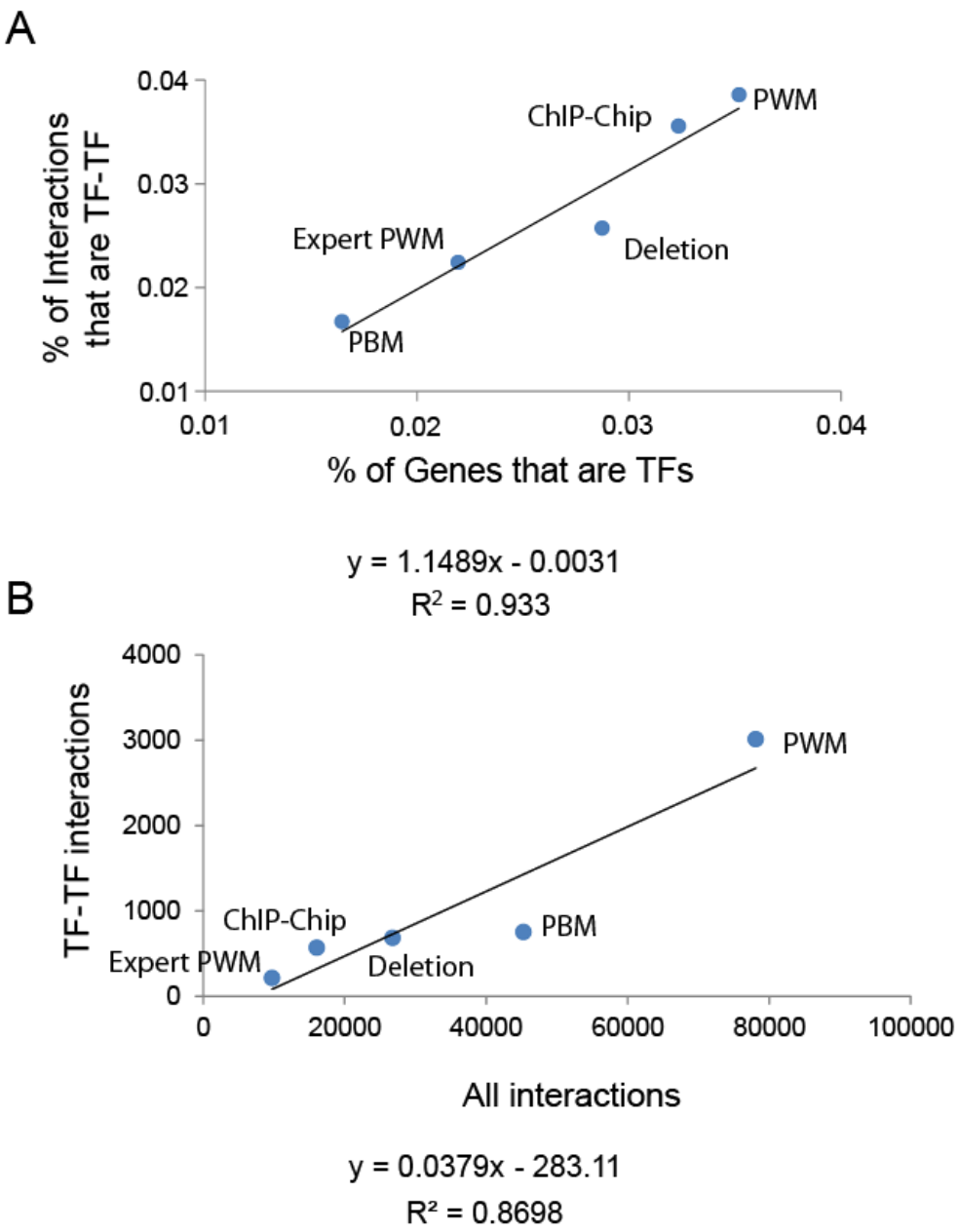
Supplemental Figure 3.2 (cont'd)

and 1. Each figure shows the mean expression of the phase cluster (dark line) and the range of values (transparent shading).

Supplemental Figure 3.3 Performance of classifier using alternative feature sets. (A) Heatmap of AUC-ROC values for SVM classification models for each cell-cycle expression set (all cell-cycle genes and genes expressed during the G1, S, S/G2, G2/M, and M/G1 phases) using TF-target interactions derived from PWM features filtered using TFs found in the ChIP-Chip data set, TFs found in the Deletion data set, and the 150 PWMs in the original PWM classifier with the highest absolute important values. The reported AUC-ROC for each classifier is the average AUC-ROC of 100 data sets composed of a balanced number of positive (cell-cycle genes) and negative (non-cell-cycle genes) classified using the parameters that maximize performance for that model (see Methods). Dark red shading indicates an AUC-ROC closer to 1

Supplemental Figure 3.3 (cont'd)

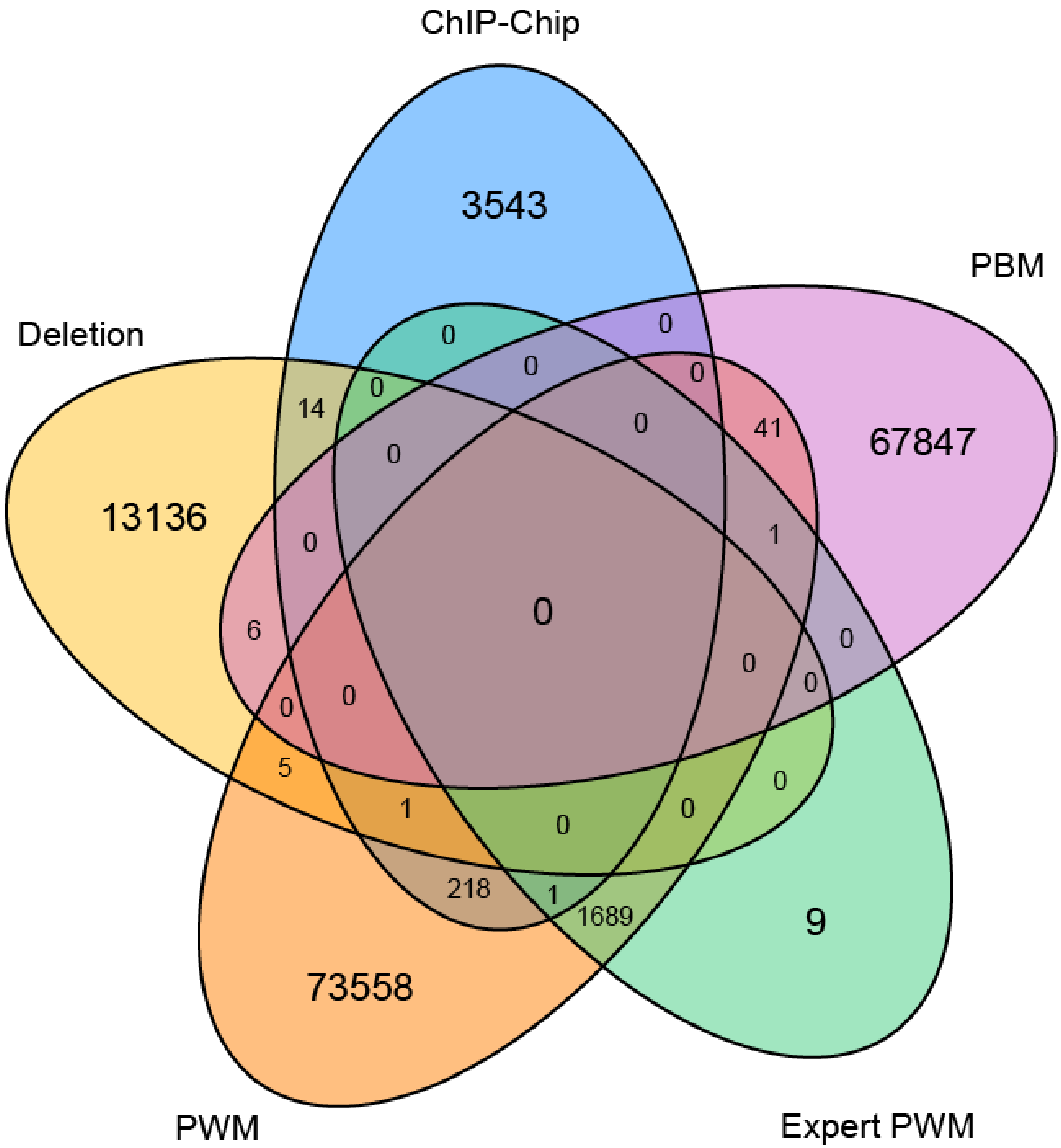
while dark blue indicates an AUC-ROC closer to zero. (B) Heatmap of AUC-ROC values for SVM classification models for each cell-cycle expression set (all cell-cycle genes and genes expressed during the G1, S, S/G2, G2/M, and M/G1 phases) using TF-target interactions derived from the ChIP-Chip, Deletion and PWM data sets filtered using the TFs covered by the PBM data set. AUR-ROC was calculated and the heatmap colored as described in (A)



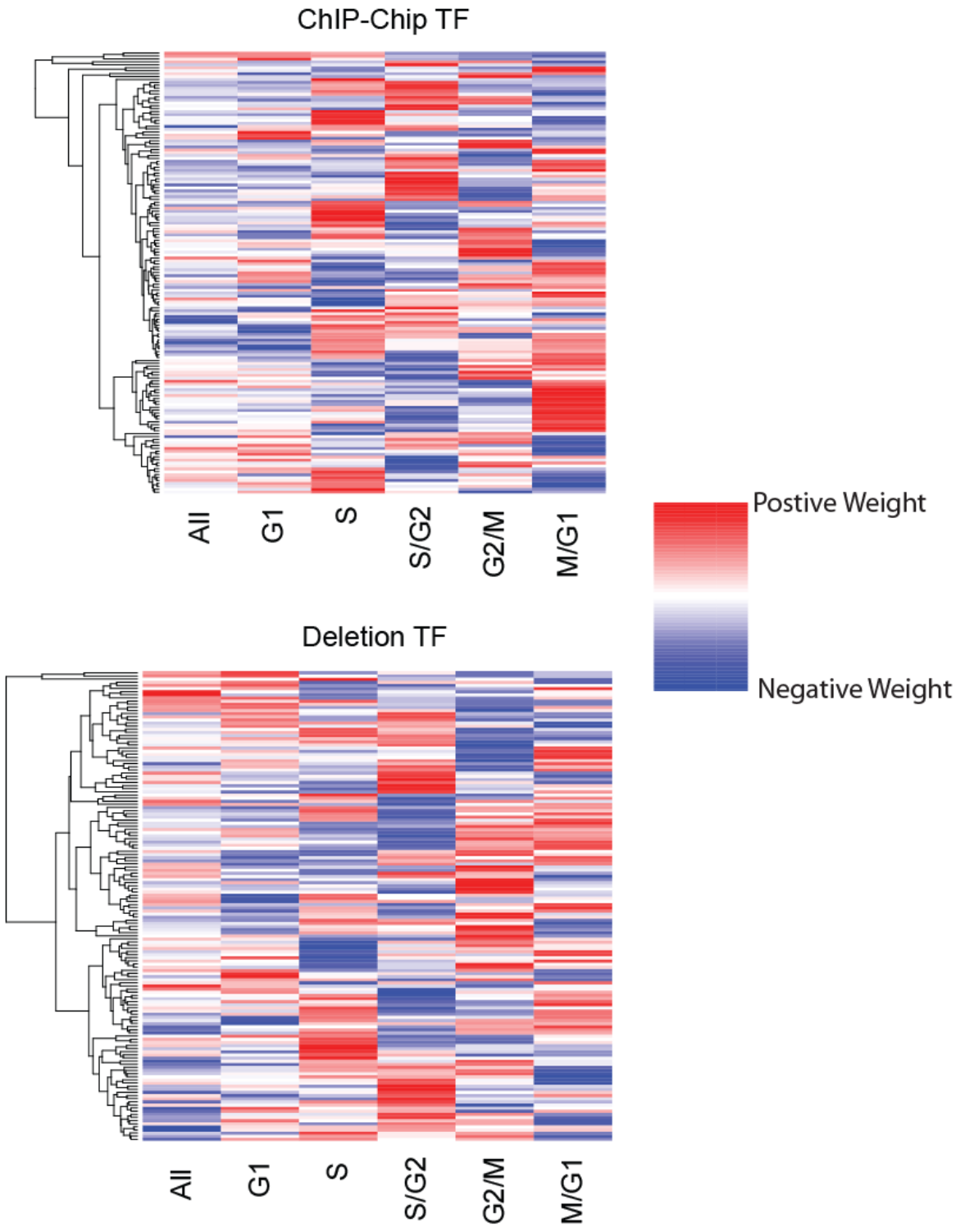
Supplemental Figure 3.4 Relationship between TF genes and TF-TF interactions. (A) The relationship between the percent of genes that are TFs (x-axis) in each of the five feature sets (ChIP-Chip, Deletion, PWM, Expert PWM, and PBM) and the percent of TF-TF interactions in that data set. Blue data points represent the observed values for each data set, and the black line is the best fit linear trendline between percent TFs and percent TF-TF interactions. The trendline equation and associated coefficient of determination are reported below the graph. (B) The

Supplemental Figure 3.4 (cont'd)

relationship between total number of interactions (x-axis) in each of the five feature sets (ChIP-Chip, Deletion, PWM, Expert PWM, and PBM) and the number of TF-TF interactions in that data set. Blue data points represent the observed values for each data set, and the black line is the best fit linear trendline between percent TFs and percent TF-TF interactions. The trendline equation and associated coefficient of determination are reported below the graph



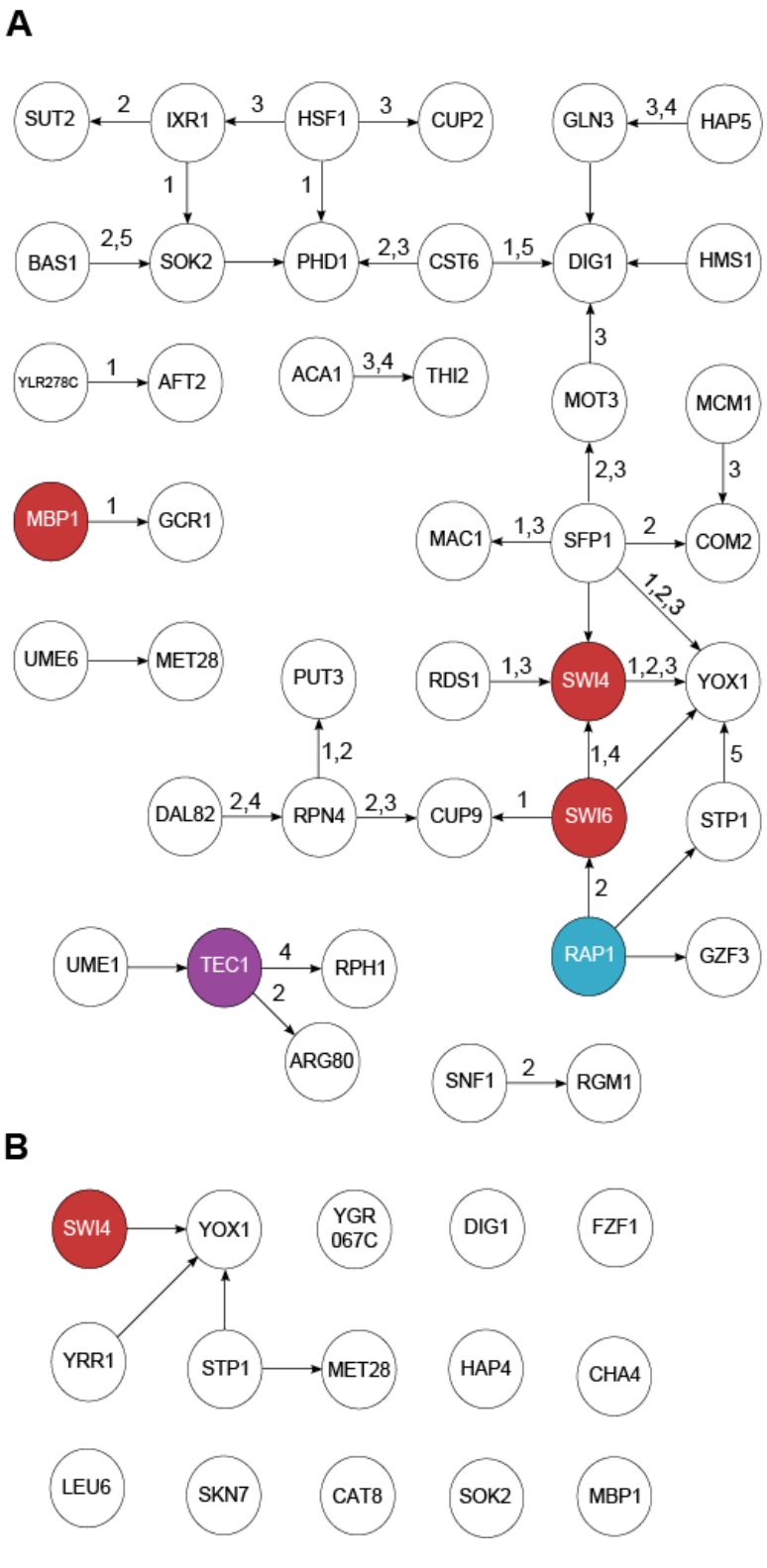
Supplemental Figure 3.5 Overlap of FFLs across data sets. Venn-diagram of the number of overlapping FFLs from different feature sets used to predict cell-cycle expression: ChIP-Chip (blue), Deletion (yellow), PWM (orange), Expert PWM (green), PBM (purple).



Supplemental Figure 3.6 Importance of TF features across classification models. Heatmaps of the importance, determined by SVM weight, of TF features from ChIP-Chip (top) and Deletion (bottom) data sets across each classifier of cell-cycle expression (All cell-cycle genes

Supplemental Figure 3.6 (cont'd)

and genes expressed during the G1, S, S/G2, G2/M and M/G1 phases). TFs in each heatmap are ordered by hierarchical clustering, and the resulting dendrogram is depicted on the left side of each heatmap. Darker red indicates that a TF has a more positive SVM weight (i.e. more enriched in cell-cycle genes) for a given model while darker blue indicates that a TF has a more negative SVM weight (i.e. more enriched in non-cell-cycle genes).



Supplemental Figure 3.7 The cell-cycle expression GRN defined using the 25th percentile of Deletion TF-TF interactions. (A) A network of TF-TF interactions selected from the Deletion

Supplemental Figure 3.7 (cont'd)

GRNs constructed using FFL features from the top 10th percentile (see Methods) of importance for predicting all cell-cycle expressed genes. Interactions are annotated with the stage of cell-cycle expression (1 = G1, 2 = S, 3 = S/G2, 4 = G2/M, 5 = M/G1) they are important for predicting (10^{th} percentile of SVM weight in Deletion models). For the purpose of contrast, elements four modules identified as being important for predicting >1 phase of cell-cycle expression in the ChIP-Chip GRN are highlighted by color: Swi6-Swi4-Mbp1 r (red), Ste12 and Tec1 (purple) and Rap1-Msn4-Hap1 (blue). This is done to illustrating the disruption of modules found in the network of TF-TF interactions selected from the ChIP-Chip GRN. (B) A network of TF-TF interactions from the Deletion GRN constructed using TF-target interactions in the top 10th percentile of importance for predicting all cell-cycle expressed genes. Genes are colored as in (A).

Supplemental Table 3.1 Coverage of cell-cycle genes by TF-target interactions in each data set

Expression Cluster	ChIP-Chip[1]	Deletion[1]	PWM[1]	Expert PWM[1]	PBM[1]
G1	238 (79%)	242 (81%)	292 (97%)	192 (64%)	250 (83%)
S	59 (83%)	62 (87%)	71 (100%)	51 (72%)	61 (86%)
S/G2	94 (78%)	107 (88%)	117 (97%)	93 (77%)	104 (86%)
G2/M1	163 (84%)	177 (91%)	190 (97%)	148 (76%)	163 (84%)
M1/G	98 (87%)	103 (91%)	112 (99%)	74 (65%)	93 (82%)
Total	652 (82%)	691 (86%)	782 (98%)	558 (70%)	671 (84%)

1. Parentheses indicate the percentage of total genes in the expression cluster covered

Supplemental Table 3.2 Coverage of cell-cycle genes by FFL interactions in each data set

Expression Cluster	ChIP-Chip[1]	Deletion[1]	PWM[1]	Expert PWM[1]	PBM[1]
G1	98 (33%)	125 (42%)	266 (89%)	40 (13%)	185 (62%)
S	20 (28%)	36 (51%)	64 (90%)	15 (21%)	51 (72%)
S/G2	34 (37%)	54 (45%)	112 (93%)	30 (25%)	83 (69%)
G2/M1	72 (39%)	104 (53%)	178 (91%)	41 (21%)	128 (66%)
M1/G	44 (39%)	66 (58%)	99 (88%)	23 (20%)	69 (61%)
Total	268 (34%)	385 (48%)	719 (90%)	149 (19%)	516 (65%)

1. Parentheses indicate the percentage of total genes in the expression cluster covered

Supplemental Table 3.3 Performance of classifiers built using TF-target interactions on only cell-cycle genes covered by ChIP-Chip FFLs

	All	G1	S	S/G2	G2/M	M/G1
AUC-ROC	0.74	0.8	0.77	0.76	0.78	0.79

Supplemental Table 3.4 Total number of feature present in each model built from combined features sets

Feature Set	Cyclic	G1	S	S/G2	G2/M	M/G1
Direct, Two-way 20%	54	84	51	52	45	41
Direct, Two-way 50%	114	113	111	114	104	93
FFLs, Two-way 20%	113	74	97	68	46	57
FFLs, Two-way 50%	263	217	221	199	126	136
Direct and FFLs, Two-way 20%	166	125	147	119	90	97
Direct and FFLs, Two-way 50%	376	329	331	312	229	228

Supplemental Table 3.5 Enrichment of TFs with cell-cycle regulation GO annotation in features of the ChIP-Chip and Deletion data sets

Feature Set	Cyclic	G1	S	S-G2	G2-M	M-G1
ChIP-Chip, 10th Percentile	7.31E-06	0.035	0.0004	0.004	0.0007	0.085
ChIP-Chip, 25th Percentile	0.0003	0.099	0.099	0.26	0.27	0.1
Deletion, 10th Percentile	0.42	0.123	0.41	0.41	1	0.11
Deletion, 25th Percentile	1	0.2755	1	0.78	0.27	0.58

Supplemental Table 3.6 Over and under enrichment of GO Terms in ChIP-Chip and Deletion feature sets

GO Term	Ench.[1]	CC[2], 10th	CC[2], 25th	D[3], 10th	D[3], 25th	Description
GO:0009074	Over	No	No	No	Yes	aromatic amino acid family catabolic process
GO:0036003	Over	No	No	No	Yes	positive regulation of transcription from RNA polymerase II promoter in response to stress
GO:0016458	Over	No	No	No	Yes	gene silencing
GO:1901717	Over	No	No	No	Yes	positive regulation of gamma-aminobutyric acid catabolic process
GO:0001133	Over	No	No	No	Yes	RNA polymerase II transcription factor activity, sequence-specific transcription regulatory region DNA binding
GO:0045848	Over	No	No	No	Yes	positive regulation of nitrogen utilization
GO:0061414	Over	No	No	No	Yes	positive regulation of transcription from RNA polymerase II promoter by a nonfermentable carbon source
GO:0061400	Over	No	No	No	Yes	positive regulation of transcription from RNA polymerase II promoter in response to calcium ion
GO:1901714	Over	No	No	No	Yes	positive regulation of urea catabolic process
GO:0008301	Over	No	No	No	Yes	DNA binding, bending
GO:0050801	Over	No	No	No	Yes	ion homeostasis
GO:0001185	Over	No	No	No	Yes	termination of RNA polymerase I transcription from promoter for nuclear large rRNA transcript
GO:0000183	Over	No	No	No	Yes	chromatin silencing at rDNA
GO:1900008	Over	No	No	No	Yes	negative regulation of extrachromosomal rDNA circle accumulation involved in cell aging

Supplemental Table 3.6 (cont'd)

GO:0061423	Over	No	No	No	Yes	positive regulation of sodium ion transport by positive regulation of transcription from RNA polymerase II promoter
GO:0042991	Over	No	No	Yes	No	transcription factor import into nucleus
GO:0031930	Over	No	No	Yes	No	mitochondria-nucleus signaling pathway
GO:0009410	Over	No	No	Yes	No	response to xenobiotic stimulus
GO:0001228	Over	No	No	Yes	No	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding
GO:0071400	Over	No	No	Yes	No	cellular response to oleic acid
GO:0035957	Over	No	No	Yes	No	positive regulation of starch catabolic process by positive regulation of transcription from RNA polymerase II promoter
GO:1900461	Over	No	No	Yes	No	positive regulation of pseudohyphal growth by positive regulation of transcription from RNA polymerase II promoter
GO:0000165	Over	No	No	Yes	Yes	MAPK cascade
GO:0031940	Over	No	No	Yes	Yes	positive regulation of chromatin silencing at telomere
GO:0001085	Over	No	No	Yes	Yes	RNA polymerase II transcription factor binding
GO:0006357	Over	No	No	Yes	Yes	regulation of transcription from RNA polymerase II promoter
GO:0071468	Over	No	No	Yes	Yes	cellular response to acidic pH
GO:1900399	Over	No	No	Yes	Yes	positive regulation of pyrimidine nucleotide biosynthetic process
GO:0031335	Over	No	No	Yes	Yes	regulation of sulfur amino acid metabolic process

Supplemental Table 3.6 (cont'd)

GO:0006986	Over	No	Yes	No	No	response to unfolded protein
GO:0030968	Over	No	Yes	No	No	endoplasmic reticulum unfolded protein response
GO:1900079	Over	No	Yes	No	No	regulation of arginine biosynthetic process
GO:0033673	Over	No	Yes	No	No	negative regulation of kinase activity
GO:0045821	Over	No	Yes	No	No	positive regulation of glycolytic process
GO:1902352	Over	No	Yes	No	No	negative regulation of filamentous growth of a population of unicellular organisms in response to starvation by negative regulation of transcription from RNA polymerase II promoter
GO:0007124	Over	No	Yes	No	No	pseudohyphal growth
GO:0071470	Over	No	Yes	No	No	cellular response to osmotic stress
GO:0090606	Over	No	Yes	No	No	single-species surface biofilm formation
GO:0010895	Over	No	Yes	No	No	negative regulation of ergosterol biosynthetic process
GO:0001103	Over	No	Yes	No	No	RNA polymerase II repressing transcription factor binding
GO:2001278	Over	No	Yes	No	No	positive regulation of leucine biosynthetic process
GO:0071940	Over	No	Yes	No	No	fungal-type cell wall assembly
GO:0000433	Over	No	Yes	No	No	negative regulation of transcription from RNA polymerase II promoter by glucose
GO:0000430	Over	No	Yes	No	No	regulation of transcription from RNA polymerase II promoter by glucose
GO:0006525	Over	No	Yes	No	No	arginine metabolic process
GO:1900081	Over	No	Yes	No	No	regulation of arginine catabolic process
GO:0019210	Over	No	Yes	No	No	kinase inhibitor activity

Supplemental Table 3.6 (cont'd)

GO:1900464	Over	No	Yes	No	No	negative regulation of cellular hyperosmotic salinity response by negative regulation of transcription from RNA polymerase II promoter
GO:0036083	Over	No	Yes	No	Yes	positive regulation of unsaturated fatty acid biosynthetic process by positive regulation of transcription from RNA polymerase II promoter
GO:0006990	Over	No	Yes	No	Yes	positive regulation of transcription from RNA polymerase II promoter involved in unfolded protein response
GO:0003700	Over	No	Yes	No	Yes	transcription factor activity, sequence-specific DNA binding
GO:0016602	Over	No	Yes	Yes	No	CCAAT-binding factor complex
GO:0043457	Over	No	Yes	Yes	No	regulation of cellular respiration
GO:0000436	Over	No	Yes	Yes	No	carbon catabolite activation of transcription from RNA polymerase II promoter
GO:0000982	Over	No	Yes	Yes	Yes	transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding
GO:0061408	Over	No	Yes	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter in response to heat stress
GO:0000977	Over	No	Yes	Yes	Yes	RNA polymerase II regulatory region sequence-specific DNA binding
GO:0046983	Over	No	Yes	Yes	Yes	protein dimerization activity
GO:0097239	Over	No	Yes	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter in response to methylglyoxal

Supplemental Table 3.6 (cont'd)

GO:0010688	Over	Yes	No	No	No	negative regulation of ribosomal protein gene transcription from RNA polymerase II promoter
GO:0045835	Over	Yes	No	No	No	negative regulation of meiotic nuclear division
GO:0032545	Over	Yes	No	No	No	CURI complex
GO:0051038	Over	Yes	No	No	No	negative regulation of transcription involved in meiotic cell cycle
GO:0090294	Over	Yes	No	No	No	nitrogen catabolite activation of transcription
GO:0000217	Over	Yes	No	No	No	DNA secondary structure binding
GO:0071406	Over	Yes	No	No	No	cellular response to methylmercury
GO:0043631	Over	Yes	No	No	No	RNA polyadenylation
GO:0046685	Over	Yes	No	No	No	response to arsenic-containing substance
GO:1990526	Over	Yes	No	No	No	Ste12p-Dig1p-Dig2p complex
GO:1990527	Over	Yes	No	No	No	Tec1p-Ste12p-Dig1p complex
GO:0001046	Over	Yes	No	No	No	core promoter sequence-specific DNA binding
GO:2000221	Over	Yes	No	No	No	negative regulation of pseudohyphal growth
GO:0061402	Over	Yes	No	No	Yes	positive regulation of transcription from RNA polymerase II promoter in response to acidic pH
GO:0001080	Over	Yes	No	No	Yes	nitrogen catabolite activation of transcription from RNA polymerase II promoter
GO:0090180	Over	Yes	No	Yes	No	positive regulation of thiamine biosynthetic process
GO:0061410	Over	Yes	No	Yes	No	positive regulation of transcription from RNA polymerase II promoter in response to ethanol
GO:0061411	Over	Yes	No	Yes	No	positive regulation of transcription from RNA polymerase II promoter in response to cold

Supplemental Table 3.6 (cont'd)

GO:0061401	Over	Yes	No	Yes	No	positive regulation of transcription from RNA polymerase II promoter in response to a hypotonic environment
GO:0061407	Over	Yes	No	Yes	No	positive regulation of transcription from RNA polymerase II promoter in response to hydrogen peroxide
GO:0001324	Over	Yes	No	Yes	No	age-dependent response to oxidative stress involved in chronological cell aging
GO:0097236	Over	Yes	No	Yes	No	positive regulation of transcription from RNA polymerase II promoter in response to zinc ion starvation
GO:0061412	Over	Yes	No	Yes	No	positive regulation of transcription from RNA polymerase II promoter in response to amino acid starvation
GO:0061422	Over	Yes	No	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter in response to alkaline pH
GO:0061429	Over	Yes	No	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter by oleic acid
GO:0005667	Over	Yes	No	Yes	Yes	transcription factor complex
GO:0032000	Over	Yes	No	Yes	Yes	positive regulation of fatty acid beta-oxidation
GO:0030154	Over	Yes	No	Yes	Yes	cell differentiation
GO:0089716	Over	Yes	No	Yes	Yes	Pip2-Oaf1 complex
GO:0001078	Over	Yes	Yes	No	No	transcriptional repressor activity, RNA polymerase II core promoter proximal region sequence-specific binding

Supplemental Table 3.6 (cont'd)

GO:0061395	Over	Yes	Yes	No	No	positive regulation of transcription from RNA polymerase II promoter in response to arsenic-containing substance
GO:0061426	Over	Yes	Yes	No	No	positive regulation of sulfite transport by positive regulation of transcription from RNA polymerase II promoter
GO:0005641	Over	Yes	Yes	No	No	nuclear envelope lumen
GO:1900436	Over	Yes	Yes	No	No	positive regulation of filamentous growth of a population of unicellular organisms in response to starvation
GO:2000218	Over	Yes	Yes	No	No	negative regulation of invasive growth in response to glucose limitation
GO:0001225	Over	Yes	Yes	No	No	RNA polymerase II transcription coactivator binding
GO:0001226	Over	Yes	Yes	No	No	RNA polymerase II transcription corepressor binding
GO:0097201	Over	Yes	Yes	No	No	negative regulation of transcription from RNA polymerase II promoter in response to stress
GO:1900240	Over	Yes	Yes	No	No	negative regulation of phenotypic switching
GO:0060963	Over	Yes	Yes	No	No	positive regulation of ribosomal protein gene transcription from RNA polymerase II promoter
GO:0071931	Over	Yes	Yes	No	No	positive regulation of transcription involved in G1/S transition of mitotic cell cycle
GO:0001076	Over	Yes	Yes	No	Yes	transcription factor activity, RNA polymerase II transcription factor binding
GO:0000790	Over	Yes	Yes	No	Yes	nuclear chromatin

Supplemental Table 3.6 (cont'd)

GO:0003676	Over	Yes	Yes	No	Yes	nucleic acid binding
GO:0071483	Over	Yes	Yes	No	Yes	cellular response to blue light
GO:0000122	Over	Yes	Yes	No	Yes	negative regulation of transcription from RNA polymerase II promoter
GO:0036095	Over	Yes	Yes	Yes	No	positive regulation of invasive growth in response to glucose limitation by positive regulation of transcription from RNA polymerase II promoter
GO:0001077	Over	Yes	Yes	Yes	Yes	transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding
GO:0008270	Over	Yes	Yes	Yes	Yes	zinc ion binding
GO:0061434	Over	Yes	Yes	Yes	Yes	regulation of replicative cell aging by regulation of transcription from RNA polymerase II promoter in response to caloric restriction
GO:0046872	Over	Yes	Yes	Yes	Yes	metal ion binding
GO:0000981	Over	Yes	Yes	Yes	Yes	RNA polymerase II transcription factor activity, sequence-specific DNA binding
GO:0003677	Over	Yes	Yes	Yes	Yes	DNA binding
GO:0006366	Over	Yes	Yes	Yes	Yes	transcription from RNA polymerase II promoter
GO:0000987	Over	Yes	Yes	Yes	Yes	core promoter proximal region sequence-specific DNA binding
GO:0061409	Over	Yes	Yes	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter in response to freezing
GO:0061403	Over	Yes	Yes	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter in response to nitrosative stress

Supplemental Table 3.6 (cont'd)

GO:0061406	Over	Yes	Yes	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter in response to glucose starvation
GO:0061405	Over	Yes	Yes	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter in response to hydrostatic pressure
GO:0061404	Over	Yes	Yes	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter in response to increased salt
GO:0006355	Over	Yes	Yes	Yes	Yes	regulation of transcription, DNA-templated
GO:0043565	Over	Yes	Yes	Yes	Yes	sequence-specific DNA binding
GO:0045944	Over	Yes	Yes	Yes	Yes	positive regulation of transcription from RNA polymerase II promoter
GO:0006351	Over	Yes	Yes	Yes	Yes	transcription, DNA-templated
GO:0000978	Over	Yes	Yes	Yes	Yes	RNA polymerase II core promoter proximal region sequence-specific DNA binding
GO:0005524	Under	No	Yes	No	No	ATP binding
GO:0016021	Under	No	Yes	No	Yes	integral component of membrane
GO:0016020	Under	No	Yes	No	Yes	membrane
GO:0005737	Under	No	Yes	No	Yes	cytoplasm
GO:0005739	Under	Yes	Yes	Yes	Yes	mitochondrion

1. Direction of enrichment
2. CC = ChIP-Chip
3. D = Deletion

Supplemental Table 3.7 Over enrichment of GO Terms in ChIP-Chip and Deletion feature sets for specific phases of cell cycle expression

Term	ChIP-Chip, 10th Percentile	ChIP-Chip, 25th Percentile	Deletion, 10th Percentile	Deletion, 25th Percentile	Unique[1]
GO:0071475	G1	NA	G1	NA	G1
GO:0006363	NA	NA	G1	NA	G1
GO:0061426	G1	G1	NA	NA	G1
GO:0031065	G1	NA	NA	NA	G1
GO:0071280	G1	NA	NA	NA	G1
GO:0045732	G1	NA	NA	NA	G1
GO:0001202	G1	NA	NA	NA	G1
GO:0005635	NA	G1	NA	NA	G1
GO:0003682	NA	G1	NA	NA	G1
GO:0072715	G2M	G2M	G2M	G2M	G2M
GO:0036086	G2M	NA	G2M	G2M	G2M
GO:0043388	NA	NA	G2M	G2M	G2M
GO:2000185	G2M	NA	G2M	NA	G2M
GO:0032048	G2M	NA	G2M	NA	G2M
GO:0000435	NA	NA	G2M	NA	G2M
GO:0033309	NA	G2M	NA	G2M	G2M
GO:0042538	NA	NA	NA	G2M	G2M
GO:0001185	NA	NA	NA	G2M	G2M
GO:0071483	NA	NA	NA	G2M	G2M
GO:0010845	NA	NA	NA	G2M	G2M
GO:1900008	NA	NA	NA	G2M	G2M
GO:0051300	NA	NA	NA	G2M	G2M
GO:0005641	G2M	NA	NA	NA	G2M

Supplemental Table 3.7 (cont'd)

GO:0000217	G2M	NA	NA	NA	G2M
GO:0043631	G2M	NA	NA	NA	G2M
GO:0061393	NA	G2M	NA	NA	G2M
GO:0046686	MG1	NA	MG1	NA	MG1
GO:0043433	NA	NA	MG1	NA	MG1
GO:0071276	NA	NA	MG1	NA	MG1
GO:0051457	NA	NA	MG1	NA	MG1
GO:1901717	MG1	MG1	NA	MG1	MG1
GO:1901714	MG1	MG1	NA	MG1	MG1
GO:0045848	MG1	MG1	NA	MG1	MG1
GO:0061415	NA	NA	NA	MG1	MG1
GO:0036003	NA	NA	NA	MG1	MG1
GO:0071466	NA	NA	NA	MG1	MG1
GO:0035948	NA	NA	NA	MG1	MG1
GO:1900079	MG1	MG1	NA	NA	MG1
GO:0034644	MG1	MG1	NA	NA	MG1
GO:0045471	MG1	MG1	NA	NA	MG1
GO:0010768	MG1	MG1	NA	NA	MG1
GO:0006525	MG1	MG1	NA	NA	MG1
GO:0000430	MG1	MG1	NA	NA	MG1
GO:1900081	MG1	MG1	NA	NA	MG1
GO:0031335	MG1	MG1	NA	NA	MG1
GO:0010038	MG1	NA	NA	NA	MG1
GO:0001128	MG1	NA	NA	NA	MG1
GO:0019413	MG1	NA	NA	NA	MG1
GO:0070211	NA	MG1	NA	NA	MG1

Supplemental Table 3.7 (cont'd)

GO:0090282	NA	MG1	NA	NA	MG1
GO:0008652	NA	MG1	NA	NA	MG1
GO:0007624	NA	NA	S	S	S
GO:0010512	NA	NA	S	S	S
	NA	S	NA	NA	S
GO:0003713	NA	S	NA	NA	S
GO:0071072	NA	S	NA	NA	S
GO:0009062	NA	S	NA	NA	S
GO:0070544	NA	S	NA	NA	S
GO:0035952	NA	S	NA	NA	S
GO:0046020	S	NA	NA	NA	S
GO:0030447	S	NA	NA	NA	S
GO:1900375	NA	NA	NA	S	S
GO:0000156	NA	NA	NA	S	S
GO:0071454	NA	NA	NA	S	S
GO:0071469	NA	NA	NA	S	S
GO:0007126	NA	NA	NA	S	S
GO:0001198	NA	NA	NA	S	S
GO:1900466	NA	NA	NA	S	S
GO:0046685	NA	NA	S	NA	S
GO:0031936	SG2	SG2	NA	NA	SG2
GO:0061425	SG2	SG2	NA	NA	SG2
GO:0000228	SG2	SG2	NA	NA	SG2
GO:0001094	SG2	SG2	NA	NA	SG2
GO:0001093	SG2	SG2	NA	NA	SG2
GO:0030466	SG2	SG2	NA	NA	SG2

Supplemental Table 3.7 (cont'd)

GO:0097235	SG2	SG2	NA	NA	SG2
GO:0061424	SG2	SG2	NA	NA	SG2
GO:0010944	NA	SG2	NA	NA	SG2
GO:0032436	SG2	NA	NA	NA	SG2
GO:0006282	SG2	NA	NA	NA	SG2
GO:0001132	SG2	NA	NA	NA	SG2
GO:0010833	SG2	NA	NA	NA	SG2
GO:0071169	SG2	NA	NA	NA	SG2
GO:0031492	SG2	NA	NA	NA	SG2
GO:0051880	SG2	NA	NA	NA	SG2
GO:0071930	NA	NA	NA	SG2	SG2
GO:0061407	NA	NA	SG2	NA	SG2
GO:0009083	NA	NA	SG2	NA	SG2
GO:0061412	NA	NA	SG2	NA	SG2
GO:0001324	NA	NA	SG2	NA	SG2
GO:0043618	NA	NA	SG2	NA	SG2
GO:0071244	NA	NA	SG2	NA	SG2
GO:0006560	NA	NA	SG2	NA	SG2
GO:0061410	SG2	NA	SG2	NA	SG2
GO:1900464	G1	G1,S	G1	G1,S	NA
GO:0061435	G1,G2M,MG1,S	G1,G2M,MG1,S	G2M	G2M	NA
GO:0036083	G2M,MG1,SG2	G2M,MG1,SG2	G2M,SG2	G2M	NA
GO:0001077	G1,G2M,S,SG2	G1,G2M,MG1,S ,SG2	G1,G2M,MG1,S ,SG2	G1,G2M,MG1,S ,SG2	NA
GO:0006366	G1,G2M,S,SG2	G1,G2M,MG1,S ,SG2	G1,G2M,MG1,S ,SG2	G1,G2M,MG1,S ,SG2	NA

Supplemental Table 3.7 (cont'd)

GO:0001080	G1,MG1	G1,MG1,SG2	MG1,SG2	G1,G2M,MG1,SG2	NA
GO:0046872	G1,MG1	G1,G2M,MG1,S,SG2	G1,G2M,MG1,SG2	G1,G2M,MG1,S,SG2	NA
GO:0003700	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0003677	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0000978	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0000981	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0006355	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0043565	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0045944	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0006351	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0060196	S,SG2	NA	G1	G1	NA
GO:0000977	NA	G1,G2M,SG2	G1	G1,G2M,MG1,SG2	NA
GO:0001078	G2M,MG1	G2M,MG1,S	G1	G1,S	NA
GO:0003676	G2M,MG1	G1,G2M,S	G1	G1,S,SG2	NA
GO:0089716	SG2	NA	G1,G2M	G1	NA

Supplemental Table 3.7 (cont'd)

GO:0031930	NA	NA	G1,MG1	G1	NA
GO:0010674	NA	NA	G1,MG1	G1,G2M	NA
GO:0061402	NA	NA	G1,SG2	G1,G2M	NA
GO:0009074	NA	G1,MG1	G1,SG2	G1,G2M,S	NA
GO:0031940	NA	NA	G1,G2M	G1,G2M	NA
GO:0032000	SG2	NA	G1,G2M	G1,G2M	NA
GO:0035969	S	S	G1,G2M,MG1	G1,G2M,MG1	NA
GO:1902352	S	S,SG2	G1,G2M,MG1	G1,G2M,MG1	NA
GO:0008270	NA	G1,G2M,MG1,S,SG2	G1,G2M,MG1,SG2	G1,G2M,MG1,S,SG2	NA
GO:1900475	NA	G2M	G1,MG1	G1,G2M,MG1	NA
GO:1900476	NA	G2M	G1,MG1	G1,G2M,MG1	NA
GO:1900472	NA	G2M	G1,MG1	G1,G2M,MG1	NA
GO:0009847	NA	G2M	G1,MG1	G1,G2M,MG1	NA
GO:1900471	NA	G2M	G1,MG1	G1,G2M,MG1	NA
GO:0001081	NA	G2M	G1,MG1	G1,G2M,MG1	NA
GO:1900525	NA	G2M	G1,MG1	G1,G2M,MG1	NA
GO:0001103	S	G1,G2M,S	G1,MG1	G1,G2M,MG1,S,SG2	NA
GO:0046983	G2M	G1,G2M,MG1,S,SG2	G1,S,SG2	G1,G2M,MG1,S,SG2	NA
GO:0097236	NA	NA	G1,SG2	G1,G2M,S,SG2	NA
GO:0097239	NA	G1,G2M	G1,SG2	G1,SG2	NA
GO:0061404	NA	NA	G1,SG2	G1,SG2	NA
GO:1902353	S	S	G1,G2M,MG1	G2M	NA

Supplemental Table 3.7 (cont'd)

GO:0005739	MG1	G1,G2M,MG1,S,SG2	G2M	G1,G2M,S,SG2	NA
GO:0034728	G1,G2M,MG1,S	MG1,S	G2M	G2M	NA
GO:2000679	NA	MG1	G2M	G2M	NA
GO:0010723	NA	MG1,S	G2M	G2M	NA
GO:0046324	NA	SG2	G2M	G2M	NA
GO:0001085	NA	G1,SG2	G2M,MG1	G1,G2M,MG1,SG2	NA
GO:0010527	NA	NA	G2M,S	G2M,S	NA
GO:0045937	NA	S	G2M,S,SG2	G1,G2M,MG1,S	NA
GO:0070210	NA	NA	MG1	G1,MG1	NA
GO:0033673	NA	G1,SG2	MG1	MG1	NA
GO:0019210	NA	G1,SG2	MG1	MG1	NA
GO:0000989	G2M	G2M	G1,MG1,S	G2M,MG1,S	NA
GO:0016458	NA	NA	MG1,SG2	MG1,SG2	NA
GO:0009410	NA	NA	G2M,MG1,S	S	NA
GO:1900399	NA	MG1	S	G2M,MG1,S	NA
GO:0043619	MG1	MG1	S	S	NA
GO:0090575	SG2	S	S,SG2	S	NA
GO:0036095	G2M,MG1	G2M	G1,SG2	SG2	NA
GO:0001010	S,SG2	S	G2M,SG2	SG2	NA
GO:0005737	G1	G1,G2M,MG1,S,SG2	SG2	G1,G2M,S,SG2	NA
GO:0061408	NA	NA	SG2	G1,G2M,SG2	NA
GO:1900240	G2M,S	G2M,S	SG2	SG2	NA
GO:2000221	MG1,S	NA	SG2	SG2	NA

Supplemental Table 3.7 (cont'd)

GO:2001158	NA	G1,G2M	SG2	SG2	NA
GO:0007124	NA	NA	G1	G2M,S	NA
GO:0000165	G2M	NA	G1	NA	NA
GO:1900463	G1	S	G1	S	NA
GO:0090606	G1	S	G1	S	NA
GO:0072363	G1,G2M,MG1,S,SG2	G1,S	G1,G2M	NA	NA
GO:0010673	NA	NA	G1,MG1	G2M	NA
GO:0007070	NA	NA	G1,MG1	NA	NA
GO:0042991	NA	NA	G1,MG1	NA	NA
GO:0070491	NA	NA	G1,MG1,S	G2M	NA
GO:1900460	G1	NA	G1,MG1,SG2	NA	NA
GO:2000218	G1,G2M,S	S	G1,SG2	NA	NA
GO:0034225	NA	NA	G1,SG2	NA	NA
GO:0035957	NA	NA	G1,SG2	NA	NA
GO:1900461	NA	NA	G1,SG2	NA	NA
GO:0090180	NA	NA	G2M,MG1,SG2	NA	NA
GO:2000222	S,SG2	NA	G2M,S	S	NA
GO:0016036	NA	NA	G2M,S,SG2	NA	NA
GO:0070417	G2M,MG1,SG2	NA	G2M,SG2	NA	NA
GO:0009450	NA	NA	G2M,SG2	NA	NA
GO:0019740	NA	NA	G2M,SG2	NA	NA
GO:0030154	G1	G1	MG1	NA	NA
GO:0001228	G1	G1	MG1	NA	NA
GO:0090295	NA	G1	MG1	NA	NA

Supplemental Table 3.7 (cont'd)

GO:0001046	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S	MG1	S	NA
GO:0005667	G1	G1	MG1,S	G2M	NA
GO:1900462	NA	NA	MG1,SG2	NA	NA
GO:0097201	MG1,SG2	MG1,S,SG2	NA	G1	NA
GO:2001043	S,SG2	S,SG2	NA	G1	NA
GO:0008301	NA	G2M,S	NA	G1	NA
GO:0000790	G1,G2M,MG1,SG2	G1,G2M,MG1,S	NA	G1,G2M	NA
GO:0045821	NA	G1,G2M,MG1	NA	G1,G2M,MG1	NA
GO:0001135	MG1	MG1	NA	G1,G2M,MG1,S	NA
GO:0006357	NA	G1,MG1	NA	G1,G2M,MG1,S	NA
GO:0000982	MG1,S,SG2	MG1,S,SG2	NA	G1,G2M,MG1,S,SG2	NA
GO:0001076	NA	G1,G2M,MG1,SG2	NA	G1,G2M,S,SG2	NA
GO:0006990	NA	G1	NA	G1,G2M,SG2	NA
GO:0016020	NA	G1,MG1	NA	G1,MG1,S	NA
GO:0061432	NA	G2M	NA	G1,S	NA
GO:0061427	NA	G2M	NA	G1,S	NA
GO:1900478	NA	G2M	NA	G1,S	NA
GO:0000122	G1,G2M,MG1,S,SG2	G1,G2M,MG1,S,SG2	NA	G1,S,SG2	NA
GO:0000987	MG1,S,SG2	G1,MG1,S,SG2	NA	G1,S,SG2	NA
GO:0016021	NA	G1,MG1,SG2	NA	G1,S,SG2	NA
GO:0001191	G1,S	G1,MG1,S,SG2	NA	G1,SG2	NA

Supplemental Table 3.7 (cont'd)

GO:0033169	NA	S	NA	G1,SG2	NA
GO:0032454	NA	S	NA	G1,SG2	NA
GO:0030907	G1,G2M	G2M	NA	G2M	NA
GO:0071931	G1,G2M,MG1	G2M	NA	G2M	NA
GO:0007074	G2M,S	G2M,S	NA	G2M	NA
GO:0000083	G2M,S	G2M,S,SG2	NA	G2M	NA
GO:0006530	S	G2M,SG2	NA	G2M	NA
GO:0004067	NA	G2M,SG2	NA	G2M	NA
GO:0001133	S	G1,G2M,MG1,S	NA	G2M,MG1	NA
GO:0061414	NA	MG1,SG2	NA	G2M,MG1	NA
GO:0070822	G1,G2M	G2M	NA	MG1	NA
GO:0003674	NA	G1,G2M	NA	MG1	NA
GO:1900423	G2M	G2M,SG2	NA	MG1,S	NA
GO:0009063	G1	G1,MG1	NA	NA	NA
GO:0042128	G1	G1,S	NA	NA	NA
GO:0001159	G1	G1,SG2	NA	NA	NA
GO:0060963	G1,G2M,MG1,S ,SG2	G1,G2M,MG1,S	NA	NA	NA
GO:0001225	G1,G2M,MG1,S ,SG2	G1,G2M,MG1,S ,SG2	NA	NA	NA
GO:0001226	G1,G2M,MG1,S ,SG2	G1,G2M,MG1,S ,SG2	NA	NA	NA
GO:2001278	G2M	G2M,MG1	NA	NA	NA
GO:0051038	G2M,MG1	MG1	NA	NA	NA
GO:0001190	MG1,SG2	MG1	NA	NA	NA

Supplemental Table 3.7 (cont'd)

GO:0010688	G1,G2M,MG1,S,SG2	S	NA	NA	NA
GO:0031496	G2M,S	S	NA	NA	NA
GO:0001012	G2M,S	S	NA	NA	NA
GO:0036033	G2M,S	S	NA	NA	NA
GO:0051019	S	MG1,S	NA	NA	NA
GO:0031848	SG2	S,SG2	NA	NA	NA
GO:0032545	G1,G2M,MG1,S,SG2	NA	NA	NA	NA
GO:0045835	G2M,MG1	NA	NA	NA	NA
GO:0000821	MG1	S	NA	NA	NA
GO:0010895	NA	G1,G2M	NA	NA	NA
GO:0044374	NA	G1,G2M,MG1	NA	NA	NA
GO:0001084	NA	G1,MG1	NA	NA	NA
GO:0010691	NA	G1,MG1	NA	NA	NA
GO:0071322	NA	G2M,S,SG2	NA	NA	NA
GO:0061416	NA	MG1,S	NA	NA	NA
GO:0070187	SG2	S	NA	NA	NA
GO:0000433	G1,G2M,S	G1,G2M,S	NA	S	NA
GO:0000304	MG1	MG1	NA	S	NA
GO:1900436	G2M	NA	NA	S	NA
GO:0031494	NA	G1	NA	S	NA
GO:0001102	NA	G1	NA	S	NA
GO:0001197	NA	G1	NA	S	NA
GO:1900465	NA	G1	NA	S	NA
GO:0061395	NA	MG1	NA	S	NA

Supplemental Table 3.7 (cont'd)

GO:0008134	G2M,S	G2M	NA	SG2	NA
GO:0035390	SG2	S,SG2	NA	SG2	NA
GO:0070200	SG2	S,SG2	NA	SG2	NA
GO:0030968	NA	G1	NA	SG2	NA
GO:0016602	G1	G1	S	NA	NA
GO:0043457	G1	G1	S	NA	NA
GO:0000436	G1	G1,MG1	S	NA	NA
GO:0061434	NA	NA	SG2	G1,G2M,S	NA
GO:0061409	NA	NA	SG2	G1,G2M,S	NA
GO:0061403	NA	NA	SG2	G1,G2M,S	NA
GO:0061406	NA	NA	SG2	G1,G2M,S	NA
GO:0061405	NA	NA	SG2	G1,G2M,S	NA
GO:0006338	NA	NA	SG2	G1,G2M,S	NA
GO:0090419	MG1	MG1	SG2	NA	NA
GO:1903468	MG1	MG1	SG2	NA	NA
GO:1990526	G2M,S	S	SG2	NA	NA
GO:0032298	MG1	NA	SG2	NA	NA
GO:0071468	G1	NA	SG2	S	NA
GO:0006572	NA	MG1	SG2	S	NA
GO:0061422	NA	NA	SG2	S	NA
GO:0061401	NA	NA	SG2	S	NA
GO:0061411	NA	NA	SG2	S	NA
GO:1990527	G2M,S	S	G2M,S,SG2	G2M,S	NA
GO:0071400	SG2	SG2	G1,MG1,SG2	G1,SG2	NA
GO:0061429	SG2	SG2	G1,SG2	G1,SG2	NA

1. Unique indicates that a GO term is only enriched in a single phase across data sets

CHAPTER 4: EXPRESSION AND REGULATORY ASYMMETRY IS A FEATURE OF RETAINED TRANSCRIPTION FACTOR DUPLICATES[1]

[1] The work described in this chapter has been submitted for publication:

Nicholas L. Panchy, Christina B. Azodi, Eamon F. Winship, Ronan C. O'Malley, Shin-Han Shiu (2017) Expression and regulatory asymmetry is a feature of retained transcription factor duplicates. *Submitted*

ABSTRACT

Transcription factors (TFs) play a key role in regulating plant development and response to environmental stimuli. While most genes revert to single copy after whole genome duplication (WGD) event, transcription factors are retained at a significantly higher rate. To assess why TF duplicates have higher rates of retention relative to other genes, we used *Arabidopsis thaliana* as a model and established linear models with expression, sequence, and conservation features to predict the extent of duplicate retention following WGD events among TFs and 19 groups of genes with other functions. We found that TFs in particular are retained more often than would be expected based on the models. Furthermore, the evolution of TF expression patterns and cis-regulatory sites favors the partitioning of ancestral states among the resulting duplicates. However, this is not because TF duplicates tend to subfunctionalize. Instead, one "ancestral" TF duplicate retains the majority of ancestral expression and cis-regulatory sites, while the “non-ancestral” duplicate is enriched for novel regulatory sites. To investigate how this pattern of biased partitioning has evolved, we modeled the retention of ancestral expression and regulatory states in duplicate pairs using a system of differential equations. In our best models, TF duplicate pairs are preferentially maintained in a partitioned state. Our findings suggest that the TF duplicates with asymmetrically partitioned ancestral states are maintained because one copy retains ancestral functions while the other, at least in some cases, acquire novel expression pattern and/or *cis*-regulatory sites.

INTRODUCTION

Plant genomes are replete with paralogous genes derived from a variety of duplication events and mechanisms (Panchy et al., 2016). Among them, whole genome duplication (WGD) events are responsible for most extant duplicate genes (Panchy et al. 2016). Two ancient WGD events took place prior to the divergence of angiosperms (Jiao et al. 2011). Subsequently, more than a dozen WGD events have occurred across a variety of angiosperm lineages (Lyons et al. 2008; Lee et al. 2013; Myburg et al. 2014; Renny-Byfield et al. 2014; Soltis et al. 2014; Wang et al. 2014), including three in the lineage leading to *Arabidopsis thaliana* (Bowers et al. 2003). As the last known WGD event in the *Saccharomyces cerevisiae* (Wolfe and Shields 1997; Kellis et al. 2004) and human (Panopoulou et al. 2003; Dehal and Boore 2005) lineages occurred prior to the radiation of angiosperms, WGD occurs more frequently in plants relative to other eukaryotic lineages.

WGD accounts for ~90% of the expansion of TF families across plants lineages (Maere et al. 2005) and TFs are consistently enriched among WGD duplicates across divergent plant species (Lespinet et al. 2002; Shiu et al. 2005; Carretero-Paulet and Fares 2012). In addition, plant TF duplicates derived from WGD are retained at higher rates than most plant genes with other functions (Seoighe and Gehring 2004; Shiu et al. 2005). These duplicate TFs contribute significantly to plant adaption (Lehti-Shiu et al. 2016), agricultural traits (Zhang et al. 2011), and domestication (Liu et al. 2015). The expansion of several TF families coincides with major events in the evolution of plants, such as the migration to land and expansion of flowering plants (De Bodt et al. 2005, Soltis et al. 2008, Weirauch and Hughes 2011). TF duplication is also central to the evolution of flowering time (Schranz et al. 2002), floral structures (Theissen and Melzer 2007) and fruit development (Litt and Irish 2003, McCarthy et al. 2015).

Because WGD results in duplication of all genes in a genome, the differences in the degrees of expansion of different gene families (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Hanada et al. 2008; Li et al. 2016) must result from differential rates of gene retention. Previously, a collection of features including sequence properties (e.g. gene length), biochemical activities (e.g. expression level), evolutionary characteristics (e.g. substitution rates), and annotated functions have been used to assess the properties of retained duplicates in general (Jiang et al. 2013; Moghe et al. 2014). It is still unclear what properties are associated with retained TFs, how well these properties explain the differences in retention rates between TFs and other function groups, and how the retained duplicate pairs differ in their properties that may shed light on the mechanisms of their retention.

In this study, we first assessed how the retention rates of *A. thaliana* WGD duplicates differ among TFs, all other genes, and genes in each of 19 other "function groups" of similar size to TFs. Next, to identify the features contributing to the differences in the percent of retained duplicates amongst different function groups of genes, we modeled the percent of retained duplicates as a function of 34 features of genes in three broad categories (expression, sequence, and conservation) in each function group. In addition, we examined whether the correlations between a feature and retention status was consistent across function groups or if some groups, like TFs, deviated from the norm. Furthermore, to assess how the ancestral and extant functions of duplicate pairs have diverged relative to their ancestral function, we determined how gene expression and *cis*-regulatory sites of TF duplicates have likely evolved post WGD by inferring the ancestral expression and *cis*-regulatory states of extant TF duplicates. Finally, we modeled the evolution of TF WGD duplicates as a system of differential equations which tracks the change in frequency of duplicate pairs retaining the ancestral state in both, one, or neither, to

assess whether the partitioning of TF duplicates pairs is maintained by a bias against losing the ancestral state in the second duplicate copy.

RESULTS AND DICUSSION

Retention of duplicate genes in different function groups following WGD

To assess the factors contributing to the differential retention of TF duplicates from WGD events and duplicates from WGD events involved in other functions, we first quantified the degree of duplicate retention of *A. thaliana* WGD duplicates in 20 different function groups. These function groups include TFs (Jin et al. 2014) and 19 other groups defined based on Gene Ontology (GO) molecular functions (see Methods). The other functional groups were chosen based on their larger sizes for comparisons with TFs and their large differences in duplicate retention (see below). Within each function group, genes were classified as "WGD-duplicates" (both duplicate copies retained) or "WGD-singletons" (only one copy retained) depending on whether there were paralogs in corresponding duplicate blocks (Bowers et al. 2003). Because duplicate retention is expected to differ across different WGD events, duplicate pairs derived from the α, β, and γ WGD events (Bowers et al. 2003) were analyzed separately. To test the association between the duplicate retention and membership in a function group, we calculated the log odds ratio of genes having a retained WGD-duplicate derived from a specific WGD event for each function group relative to all *A. thaliana* genes (see Methods). This odds ratio is used to indicate the degree of duplicate retention. If duplicate retention of a function group is not significantly different from the rest of the genome, we would expect a log odds ratio ~ 0. A positive and a negative log odds ratio indicate that a function group contains a proportionally higher and lower number of retained WGD-duplicates compared to the genome average, respectively. Among the 20 function groups examined, the log odds ratios were highly

heterogeneous and only TFs and protein kinases had significantly higher degrees of retention than the genome average for all three WGD events (**Figure 4.1**).

Although both protein kinases and TFs have odds ratio greater than the genome average across all three WGD events, based on the confidence interval of the odds of retention (**Figure 4.1**), the log odds ratios of TF retention are even higher for the older (β and γ) duplication events than those for protein kinases, indicating that on average, the longevity of TF WGD-duplicates is higher than that of protein kinases. This observation could be an artefact of the ages of these events, as some duplicates formed due to WGDs would remain in the genome, but are defined as such because they are not located in recognizable syntenic blocks. This would be particularly problematic for γ duplicates, as there are fewer syntenic regions and they are smaller (Bower et al., 2003). To address this issue, we included *A. thaliana* paralogs that may be γ WGD duplicates based the criteria used in Panchy et al. (2017) and used their synonymous substitution rate (*Ks,* see Methods) to assess if TF retention degree for older duplicates is higher than that of protein kinases. If we were to consider putative paralogs with *Ks* around the γ event ($2.7 < Ks < 2.9$) as γ WGD duplicates, the log odds ratio of retention from the γ event would still be significant for both kinases (0.52, pv = 0.02) and TFs (1.27, pv < 2.2e-16) compared to the rest of genome and TFs are still retained more frequently from this event (0.67, pv = 0.005) (**Supplemental Figure 4.1**). In summary, TFs were retained more frequently post WGD than most other function groups irrespective of group size or the timing of the event. Compared to protein kinases, one of the largest gene families in plants (Lehti-Shiu and Shiu, 2012), that also have significant higher degree of retention in all WGD events, TFs tend to be retained from older WGDs, suggesting a higher longevity of duplicates.

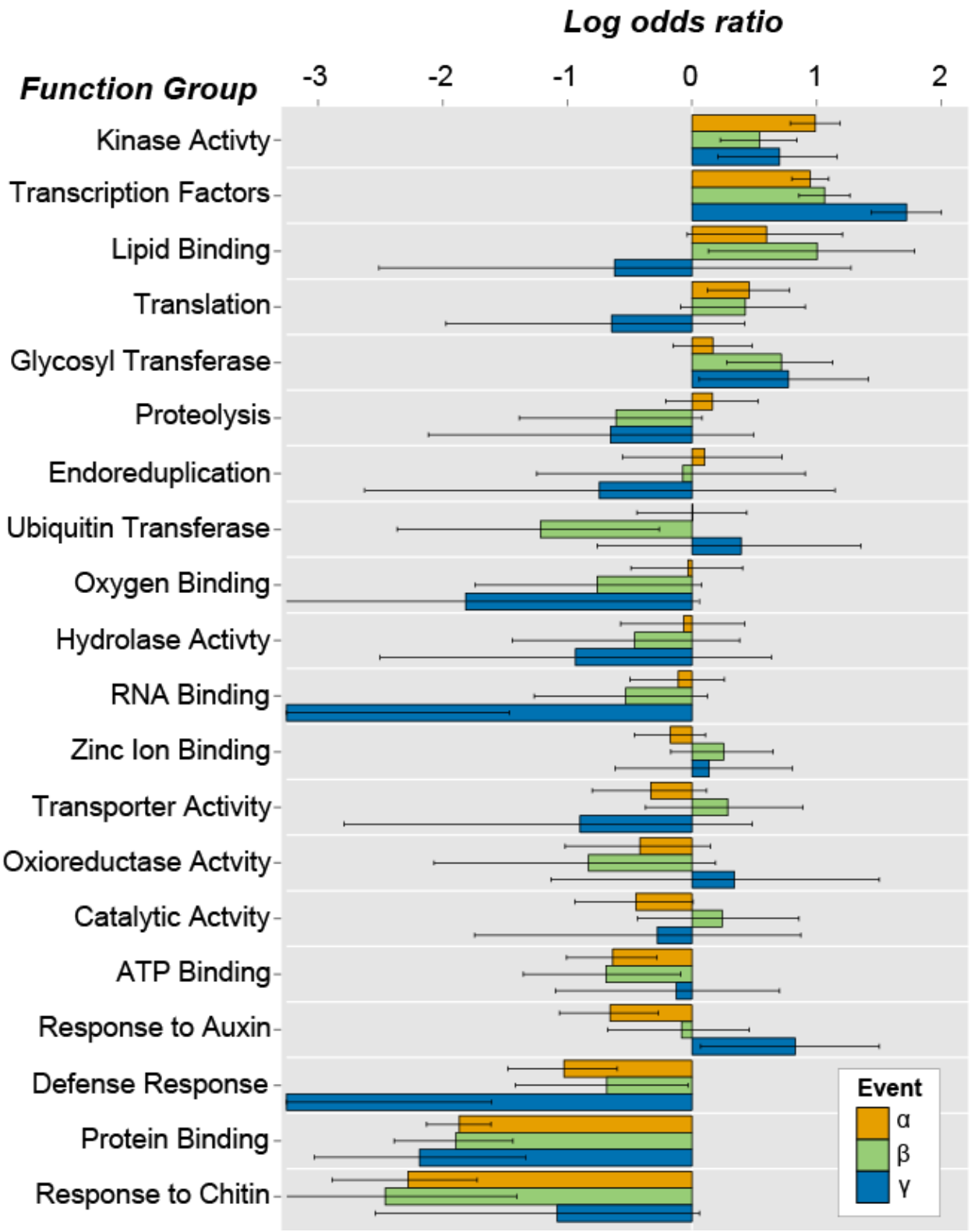


Figure 4.1 Retention of WGD-duplicate genes in A. thaliana. The duplicate gene retention rates (log odds ratios) within 20 function groups relative to whole genome. Groups are ordered

Figure 4.1 (cont'd)

by the odds in the alpha event. Colors represent different WGD duplication events (α = orange, β = green, γ = blue). Bars indicate the 95% confidence interval of the odds of retention. If the confidence interval does not overlap with zero, this indicates the odd of retaining a duplicate gene is significantly different than the genome average from that functional group at the 5%

Linear model of WGD-duplicate retention across function groups

Amongst function groups, TFs stand out as one of only two that are retained more often than the genome average consistently across all WGD events. For the rest of the function groups, the degrees of retention vary above and below the genome average across WGD events. One possibility is that the degree of retention correlates with gene numbers among functional groups. However, gene counts and degrees of retention are only very weakly correlated for any WGD event (r^2; $\alpha = 0.05$, $\beta = 0.16$, $\gamma = 0.04$; **Figure 4.2A**). Therefore, the reason for the differences in degree of retention must involve factors beyond gene function, group size, and timing of duplication. To address why the degrees of retention differs, we examined how sequence, expression, conservation, and other miscellaneous features (**Figure 4.2B, Supplemental Table 4.1**) differ among WGD-duplicate and WGD-singleton genes between function groups. We also asked how well the degree of retention differences between function groups can be explained by these different features.

To see how well the features we considered could explain the differences in degree of retention among function groups, we constructed a linear model of the degree of retention for each WGD event. Here the degree of retention is the odds ratio defined in the previous section. A subset of the features examined here has previously been shown to be significantly associated with retention of WGD-duplicates as a whole without considering individual events (Jiang et al. 2013). Here we choose to separate WGD events because there was a large variance in degree of retention across events and across function groups (**Figure 4.1**). Thus the features associated with each event and function group may differ. Consistent with this, the correlations between degrees of retention and feature values have different signs (black arrows, **Figure 4.2B**) and magnitudes (white arrows, **Figure 4.2B**) depending on the WGD events. Hence, in the next step

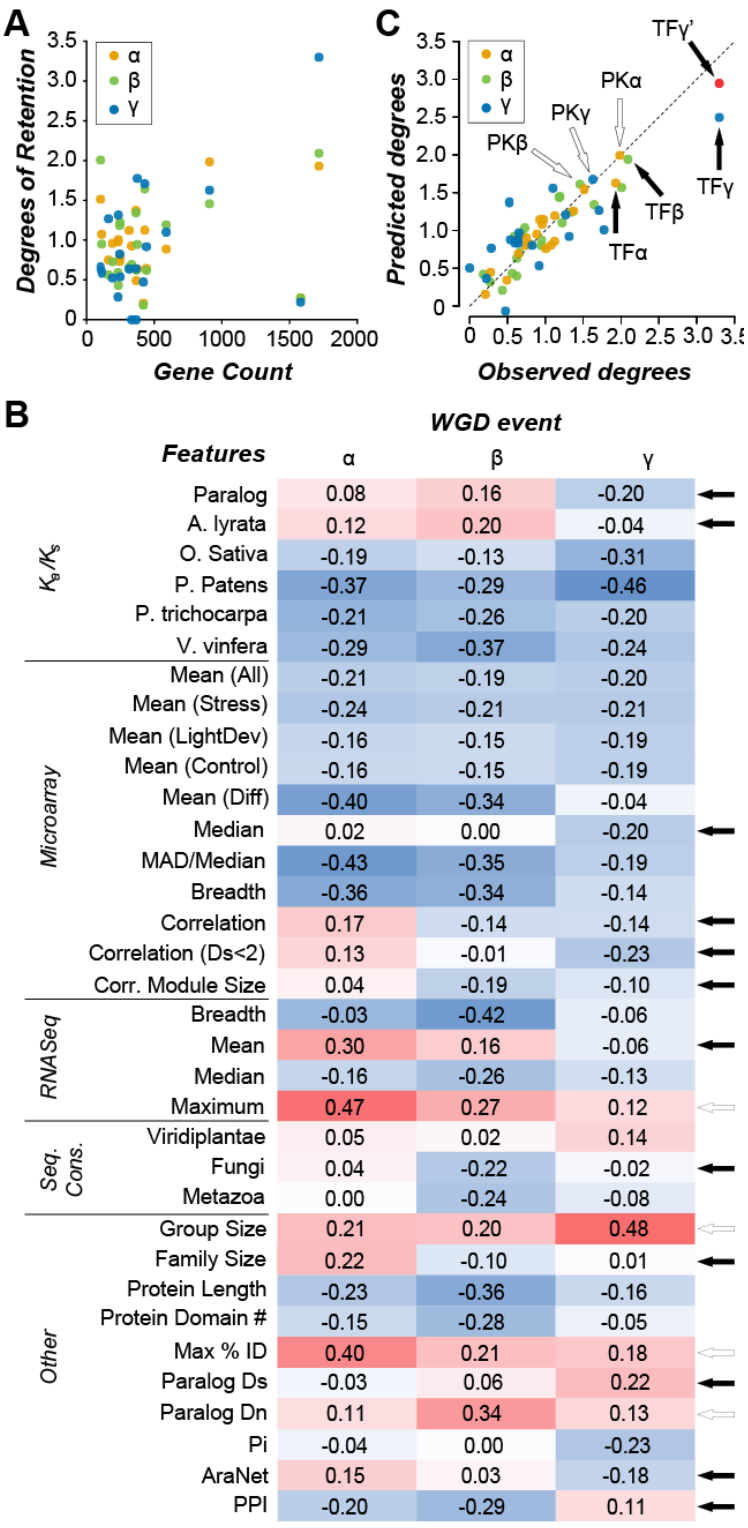


Figure 4.2 Linear model of the degree of duplicate retention in function groups based on genes features. (A) Relationships between gene counts and odds of retention of WGD duplicates

Figure 4.2 (cont'd)

across functional groups (α = orange, β = green, γ = blue). The correspondence between group sizes (numbers of genes) and degrees of retention (odds ratios) was determined using the square of the Pearson product-moment correlation coefficient (r^2, α = 0.05, β = 0.16, γ = 0.04). (B) A heatmap of the Pearson product-moment correlation coefficient (PCC) between the values of a feature across different function groups (rows) and the odds of retention of functions groups from a particular WGD event (columns, indicated by the symbols α, β, and γ). Darker red: stronger positive correlation. Darker blue: stronger negative correlation. Features with different sign of correlation across WGD events are indicated by black arrows. Features with a large (≥0.20) difference in PCCs with the same sign are indicated by open arrows. (C) The observed odds of duplicate retention (x-axis) for each group plotted against the predicted odds of retention (y-axis) from the best model for each event (α = orange, β = green, γ = blue). Dotted line: equality between predicted and observed retention odds. Values from TFs are indicated by a black arrow while values from protein kinases are indicated by an open arrow. Red dot (TFγ'): the predicted odd ratio for TFs from the γ event after adjusting for difference in percent identity of TF genes. Performance of the models was assessed by calculating the r^2 between the observed and predicted odds ratio for each event (α = 0.87, β = 0.83, γ = 0.65)

where we established linear models to predict the degree of retention with the features in **Figure 4.2C**, a model was built to describe the relationship between the average features of function groups and degree of retention for each WGD event separately. Beginning with the full set of 34 features, for each WGD event we determined the subset of features (between 5 and 6 in each case, see Methods) that maximized the F-statistic of the model (see **Supplementary File 4.1**). Our models explained 87%, 83%, and 65% of the variance in degree of retention for the α, β and γevents respectively (**Table 4.1**). Applying the F-test to the maximum F-statistic for each model, we found that each model performs significantly better at the 5% level in explaining the degrees of retention of function groups than the null model (i.e. fitting the degree of retention to their mean, **Table 4.1**).

Features explaining degrees of retention across function groups and WGD events

To determine the importance of individual features in explaining the differences in degree of retention among function groups, we determined the change in explained variance caused by independently removing each feature from the model (**Table 4.2**). Generally, degree of retention among function groups correlate with higher evolutionary rates (*Ka/Ks*) within species (paralog) but lower rate across species (orthologs to one of five species, **Figure 4.2B**). Degrees of retentions also negatively correlated with mean expression level and breadth of expression (**Figure 4.2B**). However, features with high correlation did not necessarily have a significant impact on our model performance. In contrast, the features that were retained in multiple linear models (due to their ability to maximize the F-statistic) have a greater impact on variance explained when removed. For example, maximum expression (RNA-seq) and expression mean (AtGenExpress microarray) were, respectively, positively and negatively with duplicate retention for all three WGD events. This would suggest that functional groups with genes that have more

Table 4.1 Performance of best fitting models of the odds ratio of duplicate retention

WGD Event	# Features[1]	CoD[2]	F-statistic[3]	*p*-value[4]
α	6	0.87	13.8	5.6E-05
β	5	0.83	13.2	7.1E-05
γ	5	0.65	5.1	7.2E-03

1. The number of explanatory variables (features) used in the best fitting model
2. Coefficient of Determination (r^2)
3. The F-statistic is a measure of the goodness of fit of the model to the observed odds ratio.
4. The *p*-value of goodness of fit based on the F-statistic. A significant *p*-value (< 0.05) indicates that the model performs better than fitting the mean value to the data, after accounting for the number of features in the model.

Table 4.2 Importance of features used in the linear models of duplicate retention

Feature	**Sign**[1]	α^2	β^2	γ^2
Expression Mean (AtGenExpress)	-	-0.29	-0.09	-0.49
Expression Maximum (RNASeq)	+	-0.56	-0.59	-0.14
Number of Domains	-	-0.06	-0.36	n/a
Nucleotide Diversity (Pi)	-	-0.06	n/a	-0.32
Expression Correlation (AtGenExpress)	-	n/a	-0.24	-0.21
Expression MAD/Median (AtGenExpress)	-	-0.09	n/a	n/a
Protein Length (in Amino Acids)	+	-0.07	n/a	n/a
Paralog d_n	+	n/a	-0.07	n/a
Maximum Percent Identity	+	n/a	n/a	-0.2

1. The sign of the association between the feature and duplicate retention
2. Importance of features measured as the decrease in r^2 when the feature is removed from the model, with more negative values indicating greater impact and therefore greater importance. An 'n/a' indicates the feature was not used in the model for that event.

specific expression patterns (i.e. lower average across all conditions, but higher maximum expression under a few specific conditions) tend to retain more duplicates pairs following a WGD event. Nonetheless, there are a number of cases that defy generalization due to differences across events. For example, lower expression correlation within function groups was a significant feature only in the β and γ models, while having fewer conserved domains and lower nucleotide diversity were more important to the β and γ models respectively (**Table 4.2**). These features more strongly correlated with retention of older duplicate genes suggests long term retention of duplicates favors genes experiencing stronger purifying selection (low nucleotide diversity) and those diverged expression patterns (lower expression correlation). The remaining feature were found in only one of the models and had significant but much smaller impacts the variance explained (**Table 4.2**).

Although the degree of retention predicted by the models closely align with the actual values for each function groups across each event (r^2, $\alpha = 0.87$, $\beta = 0.83$, $\gamma = 0.65$; see **Figure 4.2C**), our model based on these features alone is obviously imperfect. In particular, degree of retention for TF were consistently underestimated (black arrows, **Figure 4.2C; Supplementary Figure 4.2**), particularly in the γ model where the TF odd ratio is predicted to be only 76% of the actual value. The difficulty of predicting the degree of TF duplicate retention is likely due to the atypical feature distributions among TFs. For example, the percent base identity between a gene and its top BLAST hit within *A. thaliana* is an important predictor of the degree of retention of duplicates from the γ event. This is because the similarity of WGD-duplicates to their best BLAST for TFs (71.3% identity) resemebles the genome-wide average (72.5%), but the similarity WGD-singletons to their best hits are significantly higher (Welch's t-test, pv = 1.9e223) for TF-singletons (66.9%) compared to the genome-wide singleton average (61.3%).

This indicates that TFs, once becoming singletons, have higher degrees of sequence conservation relative to their closest paralog, presumably due to stronger selective pressure, compared to most other genes. If we inflate the mean difference in perfect identity between WGD-duplicates and WGD-singleton for TFs by a factor of 2.55 to adjust the decreased difference in percent identity, the predicted degree of TF retention of the γ event becomes 2.94 (red dot, **Figure 4.2C**), reducing the error by almost half. In addition to the linear models for predicting degrees of retention at the function group level, we have established machine learning models incorporating the same features to predict whether a gene likely has retained duplicate or not. Although this machine learning model could accurately identify genes with and without retained duplicates, the overall performance of the model and importance of features varied between kinases, TF and the rest of the genome. This suggests that, on a gene by gene level, there is dependence between gene features and functions and, therefore, these models are not useful for explain differences in the degree of retention between function groups (**Supplemental File 4.2**).

Taken together, we demonstrated that degree of retention for genes in different function groups are related to multiple features that are impacted by the timing of WGD events. However, while these features are useful for predicting the degree of retention for some function groups, they systematically underestimated degree of retention for TFs. The behavior of TFs departs from the norm in part because underlying differences in the features of TFs and genome average, suggesting their evolution following duplication likely differ significantly from other genes.

Partitioning of ancestral expression states following TF duplication

While the gene features (**Table 4.2**) were generally useful predictors of WGD-duplicates, they were less useful for predicting TF duplicates specifically. To further explore what characteristics retained TF WGD-duplicates possess, we examined how the functions of retained

TF WGD-duplicates have evolved following WGD events. Approaches to infer ancestral functions based on those of extant genes have been used to determine the rate of gene activation and repression in duplicate genes in *Drosophila melanogaster* (Oakley et al. 2006) and analyze the evolution of stress response in *A. thaliana* (Zou et al. 2009b). This approach allows for the explicit characterization of how duplicate TFs may have deviated (or not) from their ancestral state over the course of evolution, which in turn may provide information about the mechanism(s) contributing to TF retention. We first used expression patterns as a proxy of TF function(s) and inferred the likely expression states of the ancestral TFs prior to WGD (see Methods). Ancestral expression values were inferred from extant gene expression values that had been discretized into quartiles (expression state = 0, 1, 2, or 3), based on the distribution of expression levels for each expression experiment. Additionally, expression data were grouped into four subsets and analyzed separately, including light and development sets (LightDev), control conditions (Ctrl), abiotic and biotic stress treatments (Stress), and differential expression between stress treatments and controls (Diff). This grouping was then used to distinguish between trends that were universal or specific to certain datasets. We were able to infer 165,385 ancestral expression states across 474 TF WGD-duplicate pairs (a detailed breakdown of inferred states can be found in **Supplementary Table 4.2**).

To test how often the ancestral expression states of TFs are retained post-duplication, we compared the expression states of individual, extant TF WGD-duplicate to its inferred ancestral states (**Figure 4.3A**). Although all possible changes in expression state were observed between ancestral and extant TFs in each expression data subset, the most common ancestral-extant expression state combination was that the ancestral and extant TFs had the same expression quartiles (diagonal red boxes, **Figure 4.3B**). This is true across all expression quartiles, though

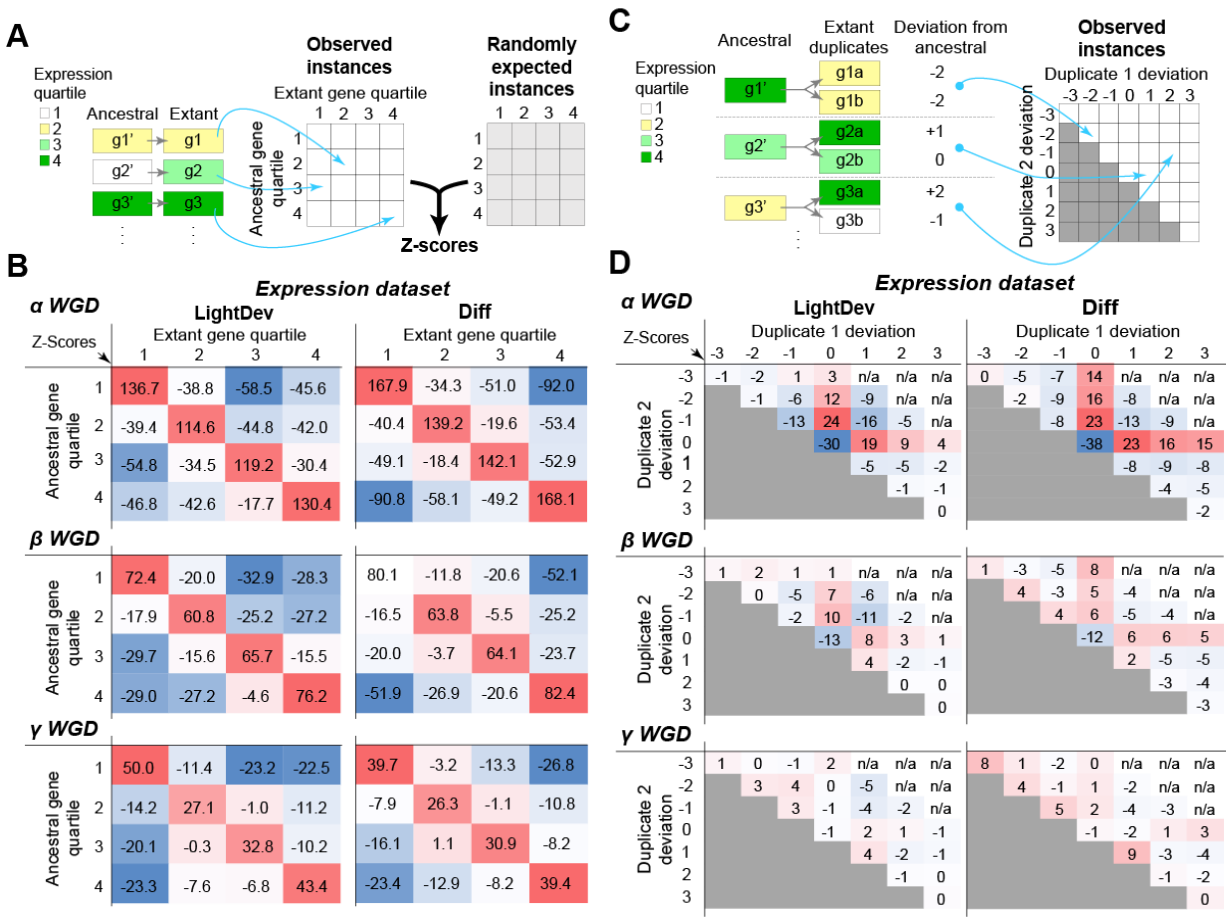


Figure 4.3 Evolution of expression in TF WGD-duplicates. (A) An illustration of how the z-scores in (B) are calculated. Individual TF duplicates are assigned to a cell using the extant (x-axis) and ancestral (y-axis) expression quartile values (dark green = 4th, green = 3rd , yellow = 2nd, white = 1st). Z-scores are then determined by comparing the frequency of the observed values to frequency distribution that would be expected if expression values were chosen randomly from a pool of extant and ancestral values. (B) Difference in expression quartile of individual TFs compared to their ancestors. Heatmaps show the z-scores of the observed frequency of each difference compared to the expected frequency for LightDev (left column) and Diff (right column) dataset in three WGD events (α = top, β = middle, γ = bottom). Darker red indicates counts further above random expectation. Darker blue indicates counts further below

Figure 4.3 (cont'd)

random expectation. (C) An illustration of how the z-scores in (D) were calculated. For each pair of WGD TF duplicates, the difference in the expression quartile values (colored the same as in (A)) of an extant duplicate and its ancestral gene is defined as "deviation". Duplicate 1 is the copy with a higher or equal expression quartile value compared to the other copy (duplicate 2). The deviation values from each duplicate copy are then used to assign the pair to a cell, where the duplicate 1 and 2 deviation values are along the x- and the y-axis, respectively. Z-scores are then determined as in (A). (D) Deviation values of pairs of TF WGD-duplicates from their ancestral state. Heatmaps show the z-scores of the observed frequency of WGD-duplicate pair deviation compared to the expected frequency for LightDev (left column) and Diff (right column) datasets in three WGD events (α = top, β = middle, γ = bottom). Color correlates with the magnitude of the z-score as in (A)

the deviation from expectation was greatest for expression values in the lowest (1) and highest (4) quartiles. This general pattern holds across all four data subsets (**Supplemental Figure 4.3**), suggesting that most TF WGD-duplicates retain their original expression irrespective of the expression context. However, when considering a pair of duplicates (**Figure 4.3C**), we found that, when the ancestral state was retained in one duplicate, it was lost more often in the other duplicate than expected by random chance (**Figure 4.3D**). This may seem to contradict the results from **Figure 4.3B**, but we should emphasize that the cases where both duplicates have the ancestral expression states are still more common (e.g. account for 53% of cases from the α-LightDev data set). However, under random permutation of duplicate pairs, 58% of α-duplicates in the LightDev data set are expected to be ancestral-ancestral (**Supplemental Table 4.3**). In contrast, we only expected 37% of pairs to be partitioned, but observed 45% pairs to have on ancestral and one non-ancestral expression state. We find the same trend for duplicates from Control, Stress, and Diff data sets originating from the relatively younger α and β events (see **Supplemental Table 4.3**).

Influence of the timing of TF duplication and expression state evolution

The "partitioned" state of TF WGD-duplicates pairs is over-represented at higher degrees for more recent WGD events (**Figure 4.3D**). In the relatively older WGD events (β and γ), having neither duplicate inherit the ancestral expression state is more common than the partitioned state where only one copy inherits the ancestral state. Using ANOVA, we confirmed that there is indeed significant interaction between the expression state of a TF WGD-duplicate pair and the timing of the WGD event ($pv < 2e\text{-}16$), which indicated that partitioning occurred relatively quickly after the most recent WGD, but that these partitioned patterns were not necessarily retained as the duplicates age.

Next we asked if TF duplicate expression levels tend to increase or decrease when they deviate away from the ancestral state. Because we found a significant interaction between the expression state evolution of TF WGD-duplicate pairs and the subset of the expression data used (pv < 3e-05), we asked this question using each expression data subset. For the LightDev (**Figure 4.3D**), Ctrl, and Stress expression subsets (**Supplemental Figure 4.4**), partitioning of ancestral expression states among duplicates favors small, negative changes from the ancestral states. Based on an earlier study showing that *A. thaliana* up-regulation in response to stress is lost more frequently that down regulation (Zou et al. 2009b), we anticipated TFs would most often lower expression quartiles compared to their ancestral state. However, when we looked at the Diff subset (the contrast between samples in the Stress subset and their respective controls) we found that TFs were equally likely to increase or decrease differential expression in response to stress compared to the ancestral state, in contrast to absolute expression levels where decrease is the norm.

To further assess how what rates of evolution are from ancestral expression states to extant states, we modeled the transition from ancestral expression (O) to higher (+) and lower (-) expression states following a WGD duplication event (see Methods). We compared a one-parameter model where the rates of transition from (O) to (+) and (-) were equal to a two-parameter model where the rates from (O) to (+) and (-) were allowed to differ (**Supplemental Figure 4.5**). The two-parameter model was significantly better than the one parameter model when absolute expression levels are considered using the LightDev (likelihood ratio test, pv = 2.2e-11), Ctrl (pv = 2.7e-3), and Stress (pv = 2.9e-3) subsets. For these subsets, the rate of evolution from (O) to (-) was 1.9~3.1 times more frequent than that from (O) to (+). For the Diff subset, (O) to (-) was only 1.2 times more frequent, which was not significant (pv = 0.43). In

summary, these results suggest that the evolution of TF duplicates favors decreasing expression levels relative to the ancestral expression state (Control, LightDev, and Stress). However, when looking at differential expression in response to stress, TF duplicates can evolve in either direction with approximately equal likelihood. Thus, following duplication, TF duplicates may have increased or decreased responses to stress, rather than losing the response altogether.

Asymmetry in the partitioning of ancestral expression and regulatory sites

Thus far we show that an ancestral expression state tends to be retained by only one copy of a TF WGD-duplicate pair and each expression state is considered individually. One outstanding question is whether each copy would retain different parts of the ancestral expression state, as would be expected if the TF duplicates were retained due to subfunctionalization (Force et al. 1999). To address this, we considered all the partitioned expression states (i.e. all expression series showing partitioning) across a pair of TF WGD-duplicates. If partitioning were random, the number of ancestral states retained by a single WGD-duplicate is expected to follow a binomial distribution for the given number of partitioned expression states and a retention probability of 0.5 (both copy equally likely to retain ancestral states). Under this scenario, the expected asymmetry of a duplicate pair (the difference in the fraction of ancestral states inherited between duplicates) is 0.18 (**Figure 4.4A**). However, the actual mean asymmetry between TF WGD-duplicates was 0.68, significantly different from random partitioning (pv < 1e-323) expected under the subfunctionalization model (**Figure 4.4B**). As with mean asymmetry, the skewed distribution of asymmetry values is also significantly different from what was expected from random partitioning (Kolmogorov–Smirnov test, pv < 2.2e-16). This biased partitioning was also found within the Ctrl (mean asymmetry= 0.84), LightDev (mean = 0.67), Stress (mean = 0.70), and Diff (mean = 0.56) subsets. To assess the

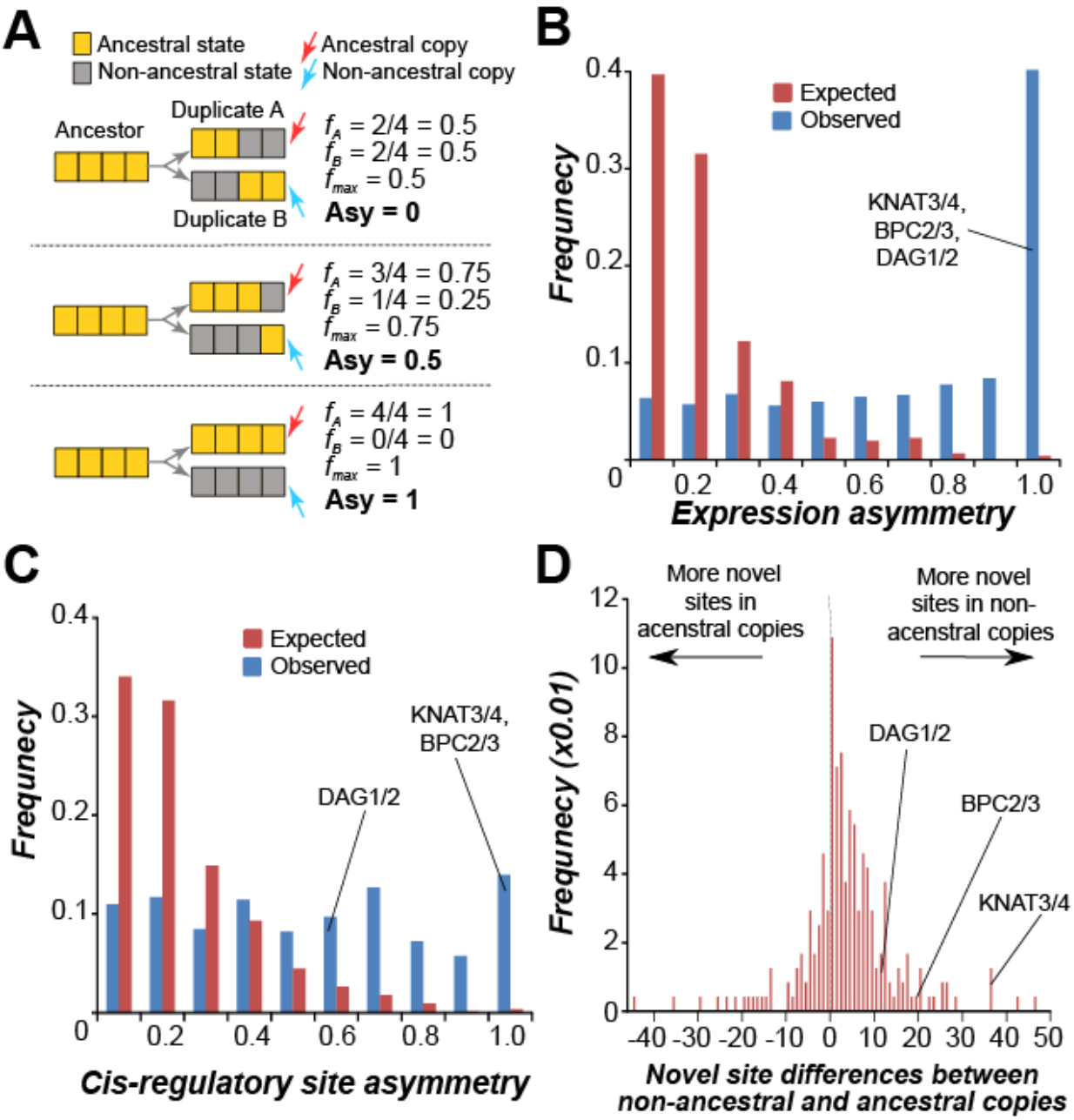


Figure 4.4 Asymmetry of ancestral state retention in TF WGD-duplicates. (A) Example of how Asymmetry score (Asy, see Methods) is calculated. Ancestral conditions are indicated by yellow boxes and non-ancestral conditions by grey boxes. Among a pair of duplicates, an 'ancestral' copy (red arrow) is the duplicate retains more ancestral states than the other, 'non-ancestral' copy (blue arrow). In case where equal numbers of ancestral states are inherited (the

Figure 4.4 (cont'd)

first case with Asy=0), the ancestral and non-ancestral designation is assigned randomly. (B) The Asymmetry scores of ancestral expression partitioning between TF WGD-duplicates. Red columns indicate the expected frequency of each score bin based on a series of grouped Bernoulli trials (see Methods) while blue columns indicated the observed frequency. (C) The Asymmetry scores of ancestral cis-regulatory site partitioning between TF WGD-duplicates. Red and blue columns are as described in (B). (D) The frequency distribution of the difference in number of novel cis¬-regulatory sites between ancestral and non-ancestral WGD duplicate copies. The value on the x-axis is calculated as the number of novel regulatory sites in the non-ancestral copy minus the number in the ancestral copy.

possibility that the observed pattern of partitioning may result from non-independent loss of ancestral expression due to the use of correlated time course data, we assembled subsets of LightDev, Stress, and Diff conditions by using only the first or last time point in each time course and found that the asymmetry scores for these subsets were virtually unchanged from those using the full datasets, the first time points (LightDev = 0.68, Stress = 0.73, Diff = 0.58) or the last (LightDev = 0.68, Stress = 0.71, Diff = 0.59) time points. Given these results, for each TF WGD-duplicate pair, we can generally define one duplicate as being "ancestral" and the other as being "non-ancestral".

Why then is the non-ancestral copy being retained? One hypothesis is that the non-ancestral copy is retained because it has acquired a novel function in the form or new expression or regulation. To test this, we first applied our model of ancestral-state partitioning to *cis*-regulatory sites. We used *cis*-regulatory sites here because the discretized expression levels used above allowed us to determine the direction of changes away from the ancestral expression state, but not whether an expression state was novel. The *cis*-regulatory sites used here are from putative binding sites of 345 *A. thaliana* TFs (O'Malley et al. 2016). We applied the same methodology used to infer ancestral gene expression to infer ancestral *cis*-regulatory sites of ancestral TFs (see Methods). In 16,015 cases, we found a *cis*-regulatory site in either one of the duplicate copies or the ancestral genes. Of these, in 57.8% of cases, an ancestral site was lost in one duplicate, while 10.5% and 16.2% cases we saw the ancestral *cis*-regulatory site lost or kept in both duplicates, respectively Similar to what we see for the partitioning of an ancestral expression state (**Figure 4.3**), loss of an ancestral *cis*-regulatory site in only one copy occurs more often than what would be expected if WGD-duplicate and ancestral genes were randomly associated (42.3%; t-test, $pv < 1e\text{-}323$). In contrast, retention (expected = 24.0%, $pv < 1e\text{-}323$)

and loss (expected = 18.5%, *pv* < 1e-323) of ancestral *cis*-regulatory sites in both WGD-duplicates were significantly less frequent than randomly expected. In addition, similar to ancestral expression state evolution, the partitioning patterns of ancestral *cis*-regulatory sites were highly asymmetric (**Figure 4.4C**), significantly different from random partitioning (Kolmogorov–Smirnov test, *pv* < 2.2e-16). Thus, much like what we observed for expression, TF WGD-duplicates can be classified into ancestral and non-ancestral copies with regard to *cis*-regulatory sites. Most importantly, amongst the 249 duplicate pairs with at least one novel regulatory site, in 71.0% of cases the non-ancestral copy had more novel *cis*-regulatory sites (**Figure 4.4C**), significantly higher than random expectation (49.8%, *pv* < 3.8e-12). Furthermore, in 61.8% of duplicate pairs the novel *cis*-regulatory sites are only found in the non-ancestral copies, compared to 14% of pairs where all of the novel sites are in the ancestral copies. These patterns suggested that, the acquisition of novel *cis*-regulatory sites likely contribute to the retention of the non-ancestral TF duplicate copies. This conclusion may be similar if we consider novel expression states, considering that the ancestral and non-ancestral designation defined according to expression data tend to have the same designation based on *cis*-regulatory sites (59.8%, compared to expected by random association at 24.6%, *pv* = 1.8e-20).

Within this pool of duplicate pairs with distinct ancestral and non-ancestral duplicates, there are a number of examples the non-ancestral copies exhibiting a different function from the ancestral duplicate. The non-ancestral gene KNAT4 has gained 37 novel regulatory sites relative to KNAT3, which has retained all partitioned expression and regulatory sites (**Figure 4.5A**). In this case, both genes retain a development regulatory function, but KNAT4 is primarily found in the elongation zone of roots in phloem and pericycle cells, while KNAT3 is found in the differentiation zone in pericycle and cortex cells (Truernit et al., 2006, Truernit and Haseloff,

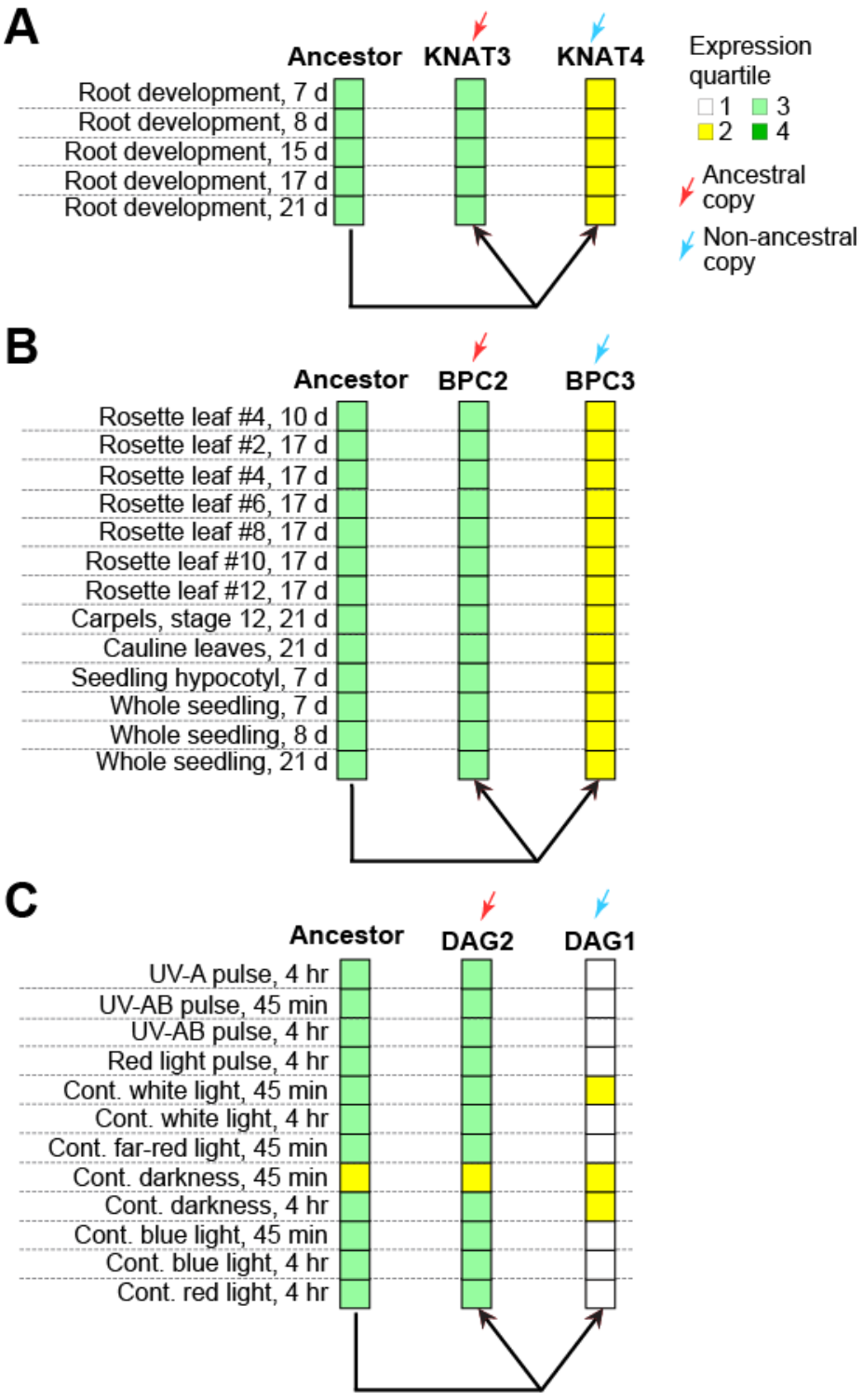


Figure 4.5 Expression partitioning between duplicate pairs with high regulatory asymmetry. Expression partitioning of three duplicate pairs KNAT3/4 (A), BCP2/3 (B), DAG1/2 (C) where the non-ancestral duplicate (blue arrow) exhibits differential function from

Figure 4.5 (cont'd)

the ancestral duplicate (red arrow). Expression quartile is indicated by color (dark green = 4th, green = 3rd, yellow = 2nd, white = 1st). Note that only expression conditions under which function differs between the duplicates are shown

2007). In another example, BPC3 is a non-ancestral duplicate which has 20 novel regulatory sites and has lost ancestral expression in 15 conditions where it is retained in its duplicate, BPC2 (**Figure 4.5B**). Previous research found BPC3 functions antagonistically not only to BPC2, but other members of BPC regulatory family in regard to controlling growth, leaf shape, and flower development (Monfared et al., 2011). Finally, the non-ancestral copy DAG2 is a positive regulator of phyB induced germination which is directly regulated by the ancestral copy DAG1 (Santopolo et al. 2015). This is of particular interest because, in spite of having opposite regulatory roles, both duplicates have similar expression breadth (Gualbertia et al 2002) and our own data indicates that ~40% of inferred ancestral *cis*-regulatory elements are conserved in both copies even though the DAG1 retains most of the ancestral response to light (**Figure 4.5C**). This indicates that function differentiation can arise even when ancestral expression is incompletely partitioned between copies.

Patterns of WGD-duplicate divergences and partitioning results from evolutionary bias

We have demonstrated that partitioning of ancestral expression and regulation into ancestral and non-ancestral duplicates is favored following duplication of TFs. It remains an open question if this ancestral state partitioning is maintained and thus the duplicate retains the ancestral expression/regulation is likely under selection. Alternatively, if the rate of ancestral state loss of the second copy is similar to that of the first, it would suggest the partitioning is simply a transition state and is not maintained. To determine which of the above cases is likely true, we modeled loss of ancestral states of TF WGD-duplicate pairs (see Methods). Using the synonymous substitution rate (d_s) of TF WGD-duplicate pairs derived from the α, β, or γ events as a proxy for time, the rate of transition between WGD-duplicate pairs where neither (state O), only one (state I), or both (state I) duplicates had lost ancestral expression was modeled (**Figure**

4.6A). We compared a model where the rates of transitions between all states were equivalent (same rates for losing the ancestral states in both duplicates, one-parameter model) with a model where the transition rates between state O and I were allowed to vary from those between state I and O (two-parameter model). These models were applied to all expression subsets (**Supplemental Figure 4.6**), the results of the one and the two parameter models using the LightDev dataset are shown as an example in **Figure 4.6B**.

We found the two-parameter model to be significantly better at explaining the observed difference in WGD-duplicate states over time (Likelihood Ratio Test, *p*-value < 2e-14). Regardless of the expression data set, the transition rates between state O (ancestral expression in both duplicates) and I (ancestral expression in on duplicate) were 7-13 times higher than the rates between state I and II (ancestral expression in neither duplicate) (**Figure 4.6**). In addition, based on the estimated rates of ancestral state loss over time by the two-parameter model (curves in **Figure 4.6B**), the number of partitioned WGD-duplicates accumulated rapidly post WGD, followed by a relatively slow accumulation of bases where ancestral expression states had been lost in both duplicates. We also assessed a four-parameter model (O→I, I→II, II→I, I→O) that was not better than the two-parameter model. Applying this same approach to model regulatory site evolution revealed that the best fit model for regulatory site evolution involved allowing all four rate parameters to vary (*p*-values 4.8e-13 and 1.2e-11 vs. one and two-parameter models, respectively; **Figure 4.6C**). The rates governing the O→I transition (*x)* are two orders of magnitude higher than the I→II transition (*w,* **Figure 4.6D**). Importantly, in the four-parameter model, there was a high rate of O→I transition estimated at the early stage of WGD (blue curve, **Figure 4.6C**). In addition, an appreciable proportion of partitioned duplicates lost ancestral regulatory sites in the second copy (green curve, **Figure 4.6C**) that contributed to the pattern

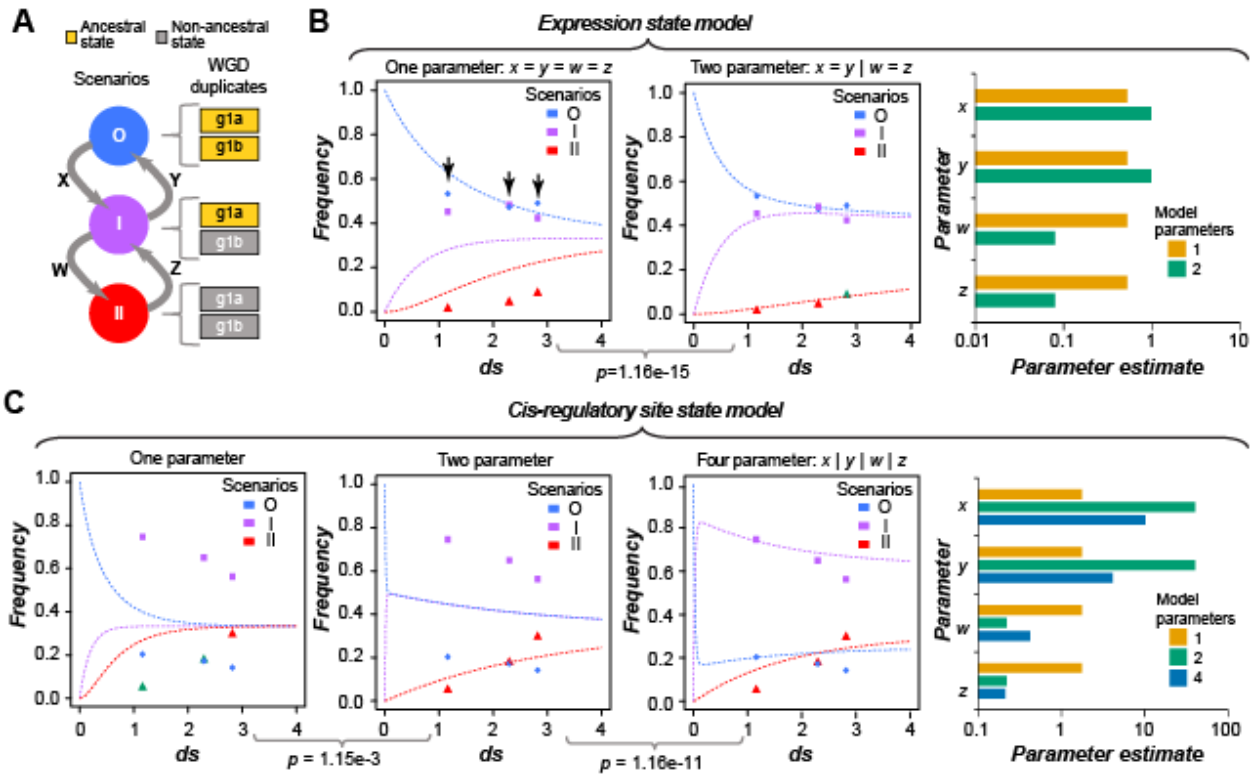


Figure 4.6 ODE models of TF WGD-duplicate expression and cis-regulatory site evolution relative to the ancestral state. (A) In this model, we consider the transition of WGD-duplicate pair expression states between three possible scenarios (O = both retained, I = one retained, II = neither retained) using four variables representing the rate of transition between state (x,y,w,z). (B) Left and middle: results for the one parameter (x=y=w=z) and two parameter (x=y|w=z) versions of the expression state model showing the change in time (x-axis) and the frequency (y-axis) of each scenarios. Curves represent the continuous output of the model in different scenarios. The significance of including additional parameters (the p-value between the curly brackets) was determined using the likelihood ratio test. Right: A bar graph of the parameter values for the one (orange) and two (green) parameter versions of the expression ODE model. (C) Left three sub-graphs: results for the one parameter (x=y=w=z), two parameter (x=y|w=z), and four parameter (x|y|w|z) versions of the cis-regulatory site model showing the change in time

Figure 4.6 (cont'd)

(x-axis) and the frequency (y-axis) of each WGD-duplicate-pair scenario. Curves represent the continuous output of the model in different scenarios. The p-values are derived from the likelihood ratio tests between models. Far right: a bar graph of the parameter values (x,y,w,z) for the one (orange), two (green), and four (blue) parameter versions for the cis-regulatory site ODE model.

where the proportion of partitioned duplicated peaked at low *ds* followed by a reduction. This is in sharp contrast compared to the transition rate estimate over time for expression where second copies tend not to lose ancestral expression state (**Figure 4.6B)**, indicating that regulatory sites are faster evolving and more labile compared to expression states.

CONCLUSIONS

In this study, we have shown that duplicates are retained at different rates across function groups. In addition, we established linear models to assess how expression, conservation, and sequence structural features of genes in these functional groups may explain their retention rate difference. Although the linear model is far from perfect, it serves as the basis for exploring more complicated interactions underlying duplicate retention, i.e., the potential interaction between gene features and annotated gene function suggested by our results. We also demonstrate a preference for maintaining partitioned expression and regulatory site states between TF WGD-duplicate pairs. Yet, while we have established that retained duplicate genes have distinct expression, sequence and regulatory features and TF duplicate genes in particular are characterized by asymmetric-partitioning, the question of what this implies about why duplicate genes are retained remains to be addressed.

Many mechanisms have been proposed to explain why duplicate genes are retained. Any duplicate pair could potentially be retained via neofunctionalization (Ohno,1970)) or escape from adaptive conflict (Des Marais and Rausher 2008) which involve the evolution of new or improved function that is positively selected for. However, subfunctionalization (Force et al. 1999) or gene balance (Birchler and Veitia 2007; Birchler and Veitia 2010; Baker et al. 2013) are specific to TFs and other gene with a large number of interactions/functions (Seoighe and Gehring 2004; Maere et al. 2005; Shiu et al. 2005, Alvarez-Ponce and Fares 2012) which all need to maintained following WGD. On an experiment-by-experiment basis, the partitioning of ancestral expression states **(Figure 4.3D)** would appear to support the notion of WGD-duplicate retention by subfunctionalization (Force et al., 1999). However, when examining the ancestral state partitioning patterns across multiple experiments, we find an extreme bias where one TF

duplicate retains most of the ancestral states and the other, non-ancestral copy retains few or none (**Figure 4.5**). Most importantly, we showed that the non-ancestral copy tends to gain novel *cis*-regulatory sites (**Figure 4.5D**) and exhibit differential expression from the ancestral state. This pattern harkens back to the notion of there being an ancestral copy and a neofunctionalized copy after duplication, contributing to the retention of both duplicates (Ohno, 1970). This would appear to be the case for duplicate pairs like KNAT3/4, BPC2/3, and DAG1/2.

Nonetheless, we should note that there remain asymmetrically partitioned duplicates that are retained without clear evidence of neofunctionalization. A clear case of this is the duplicate pair ANAC19/72 which, in spite of ANAC72 gaining novel 21 regulatory sites, appears to have redundant function regulating stress response, both with each other and with others NAC-family TFs (Tran et al., 2004; Zheng et al., 2012; Takasaki et al., 2015). It has been theorized that seemingly redundant duplicates may be retained due to subfunctionalization at the expression level following a reduction expression after duplication and/or subsequent "rebalancing" of expression that could be positively selected (Qian et al., 2011). Yet while this might explain the retention of asymmetric duplicates with similar function, it cannot explain the retention duplicate pairs where ancestral expression in maintained across both copies. For example, 7.5% and 13.9% of duplicates TF pairs have retained >80% of ancestral expression in both copies in the Stress and LightDev data set respectively. Thus, while neofunctionalization and subfunctionalization may explain the retention of partitioned duplicates, the presence of duplicate pairs with a high degree of ancestral expression in both copies and the overall prevalence of retaining ancestral expression following the α and β WGD events (**Supplemental Table 4.3**) remains to be addressed.

If subfunctionalization does not play a predominant role in TF duplicate retention, what other mechanisms may be responsible? One prominent hypothesis is gene balance where stipulate that duplicate genes with products that form multimeric complexes will tend to be retained to maintain the stoichiometry (Birchler and Veitia 2007; Birchler and Veitia 2010) and enables future sub- and/or neofunctionalization (Veitia et al. 2013). If gene balance does play a significant role in retention of TF duplicates, we would expect duplicates from more recent WGD events to have higher proportion of cases where both copies retained the ancestral expression and regulatory site states. However, our ODE model for evolution of duplicate TF pairs indicate that the proportion of duplicates both retaining ancestral states reduces quickly following WGD (**Figure 4.6**), suggesting that, if gene balance plays a significant role, it is limited to the initial period after WGD. The caveat is that our current ODE models of ancestral expression and regulatory site evolution is based on WGD events that are >50 million old. It will be useful to incorporate data from other species with more recent WGD events into the model to further address this question. Additionally, WGD-Duplicates TFs are known to be preferentially retained across many plant species (Carretero-Paulet and Fares 2012), it will be of interest to see if the patterns of ancestral expression and regulatory site partitioning we have uncovered in *A. thaliana* are shared in other plant lineages sharing the α, β, and γ events, or common to any species with ancient WGD events. Furthermore, our study focuses on the overall pattern of TF evolution. It is anticipated that different TF families will evolve differently from each other. In future studies, it will be important to directly compare the size, rate of retention, and rate of partitioning both within and across species in individual families.

MATERIALS AND METHODS

Genome sequences, gene annotation, and Expression Data

Genome sequences, protein sequences, and gene annotation information for *A. thaliana* was obtained from Phytozome v10 (https://phytozome.jgi.doe.gov/pz/portal.html). WGDs were defined according to Bowers et al. (2003) and tandem genes in *A. thaliana* were defined as pairs of reciprocal best BLAST hits with an e-value < 1e-10 and ≤ 5 intervening genes. Expression microarray data for this study was taken from AtGenExpress (Schmid et al. 2005; Kilian et al. 2007; Goda et al. 2008), normalized using RMA (Irizarry et al. 2003) in R as performed previously (Zou et al. 2009a). The array data was divided into four groups: control conditions (in environmental condition experiments, Ctrl), light and development set (LightDev), abiotic and biotic stress treatments (Stress), and differential expression between stress treatments and controls (Diff) (**Supplemental Table 4.4**). The Diff data contains the log2 normalized difference between data sets for each stress condition/treatment/duration and its corresponding controls. In addition to microarray data, we have included a set of 214 RNA-sequencing samples (**Supplemental Table 4.5**) from *A. thaliana* Col1 wildtype from the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) as of September 30, 2014. Raw sequence reads were processed using Trimmomatic (Bolger et al. 2014), with a quality threshold of 20, window size of 4, and hard-clipping length of 3 for leading and trailing bases. Processed reads were then mapped to the *A. thaliana* genome using Tophat2 (Kim et al. 2013) and expression levels calculated with Cufflinks (Trapnell et al. 2010), both with a maximum intron length of 5,000bp.

Defining TFs and other groups of genes in A. thaliana

TFs were defined according to the criteria used by the Plant Transcription Factor Database (Jin et al. 2014) with 1,717 annotated TF loci in *A. thaliana*. To assess the degrees of TF duplicate retention after each WGD event, we defined a set of "functional groups" for comparison following from the procedure used in Maere et al. (2005). To compare among genes with divergent functions and to ensure the log odds indicative of the degrees of retention could be defined for each group, function groups were defined using Gene Ontology (GO) (Ashburner et al. 2000) terms in the molecular function and biological process categories from The *Arabidopsis* Information Resource (https://www.arabidopsis.org/), and only groups containing 100-2,000 genes and ≥20 WGD-duplicate pairs were kept. We excluded GO:0006355 (regulation of transcription, DNA-templated) due to its substantial overlap with the TF group we have defined above. The remaining 19 function groups include: ATP Binding (GO:0005524), catalytic activity (GO:0003824), defense response (GO:0006952), DNA endoreduplication (GO:0042023), hydrolase activity hydrolyzing O-glycosyl compounds (GO:0004553), kinase activity (GO:0016301), lipid binding (GO:0008289), oxidoreductase activity (GO:001649), oxygen binding (GO:0019825), protein binding (GO:0005515), proteolysis (GO:0006508), response to auxin (GO:0009733), response to chitin (GO:0010200), RNA binding (GO:0003723), transferase activity, transferring glycosyl groups (GO:0016757), translation (GO:0006412), transporter activity (GO:0005215), ubiquitin-protein transferase activity (GO:0004842), zinc ion binding (GO:0008270). A list of genes in each group can be found in **Supplemental Table 4.6**.

Fitting odds ratio of duplicate retention within each group of genes for each WGD event using linear models

A gene was designated as a "WGD-duplicate" if its paralog derived from a particular WGD event is present. For a gene without its paralog from WGD, it was designated as a "WGD-singleton" gene. The degree of retention for a function group, *g*, after a specific WGD event, *w*, is defined as:

$$R_{g,w} = \frac{(D_{g,w}/S_{g,w})}{(D_{\neg g,w}/S_{\neg g,w})}$$

Where $D_{g,w}$ and $D_{\neg g,w}$ are the numbers of WGD-duplicate genes in group *g* and those not in group g (¬g), respectively. $S_{g,w}$ and $S_{\neg g,w}$ are the numbers of WGD-singleton genes in group *g* and those not in group g (¬g), respectively. The 95% confidence interval around the point-estimate $R_{g,w}$ was defined using the "fisher.exact" function in R, the details of which can be found at in Fay (2010). For each WGD event, we established a general linear model with the glm function in the R environment which relates the $R_{g,w}$ to a set of features of each gene group. The 34 features (predictor variables, **Supplemental Table 4.1**) were filtered with the following procedures to prevent over-fitting because we have only 20 function groups. We calculated the correlation between all features to find all cases where the absolute value of correlation was >0.7. The considerations for which features to keep included: (1) how well each feature correlated with $R_{g,w}$ on its own, (2) whether the feature was derived from a subset of another feature, and (3) the number of other features with a correlation > 0.7 (favored the elimination of more features). In addition to the above criteria, one data set (protein-protein interactions) was eliminated because of a high frequency of missing values (88%). The synonymous substitution rate (K_S) feature and any feature using K_S in their calculation were also excluded because they would be highly correlated with WGD timing and confound our analyses comparing the three

WGD events. The filtering step left 11 features for building the general linear model. Following fitting the glm function, features were ranked according to their *p* values from the least to the greatest and the feature with the largest p-value was dropped. The model was then fit to the reduced feature set and features were once again ranked. This process was repeated until the F-statistic (a measure of goodness of fit of the given model against a null model where all coefficients are set to zero) of the model was maximized and the final *p* value was calculated based on the maximal F-statistic. The final model for each event can be found in **Supplementary File 4.1**.

Inferring ancestral expression levels and cis-regulatory sites

DNA-binding domains were identified in TF protein coding sequences using hmmscan via HMMER3 (Mistry et al. 2013) based on the Pfam-A version 29.0 HMMs (Finn et al. 2016) with a threshold e-value of 1e-5. TFs were classified into families according to their DNA-binding domains and 44 of 59 TF families with ≥4 members were used for further analysis (**Supplemental Table 4.7**). For each TF family, full-length protein sequences were aligned using MAFFT (Katoh and Standley 2013) with default parameters. The phylogeny of each TF family was obtained using RAxML (Stamatakis 2014) with the following approach: rapid Bootstrapping algorithm, 100 runs, GAMMA rate heterogeneity, and the JTT amino-acid substitution model. These trees were then mid-point rooted with retree in PHYLIP (Felsenstein, 1989) and used to infer the ancestral gene expression states and the *cis*-regulatory sites of WGD-duplicate TF pairs with BayesTrait (Pagel et al. 2004) as was done in our earlier study (Zou et al. 2009a). The expression data sets used are described in **Supplemental Table 4.4**. The discretized gene expression state (0,1,2,3) was based on the quartiles of gene expression levels within each experiment. Thus the inferred, ancestral expression state was also discretized. For *cis*-regulatory

sites, the binding targets of 345 *A. thaliana* TFs were defined based DNA Affinity Purification-Sequencing data (O'Malley et al., 2016) from the Plant Cistrome Database (http://neomorph.salk.edu/dap_web/pages/index.php) where at least 5% of the read associated with a site were found to be in the 200bp peak region. We inferred whether a particular site was present or absent (0,1) in the common ancestor of a duplicate pair. For both expression and regulatory site data, in cases where there was a missing value, it was explicitly included as an ambiguous state. To call the ancestral state from the expression or *cis*-regulatory site data, we required a posterior probability > 0.5. Cases where the called state was ambiguous or no majority existed were excluded from further analysis.

Asymmetry of the retention of ancestral expression and regulatory sites

For determining expression state asymmetry, only TF WGD-duplicates with ≥5 partitioned ancestral expression states in one of the four expression datasets (Ctrl, LightDev, Stress, and Diff) were considered. For a WGD-duplicate pair with genes A and B, if the number of inherited ancestral expression states in A was larger or equal to that in B, then A and B were defined as the ancestral and the non-ancestral duplicate copies, respectively. The degree of asymmetry ($Y_{A,B}$) of expression states between two duplicates was defined as:

$$Y_{A,B} = \max(F_A, F_B) - (1 - \max(F_A, F_B))$$

Where F_A and F_B are the frequency with which ancestral expression was retained for duplicates A and B, respectively. By definition, $F_A + F_B = 1$, such that $Y_{A,B}$ has value between 0 (when $F_A = F_B$, no asymmetry) and 1 (when either F_A or $F_B = 1$, maximum asymmetry).

With the asymmetry values for each TF pair, an average asymmetry value of all TF pairs was calculated for each expression dataset, as well as for the union of all TF duplicates from all datasets (1,239 values total) to assess how the observed degree of asymmetry compared to what

would be expected from if every partitioned state was independent (i.e. each gene has an equal chance of retaining the ancestral state regardless of the outcome of previous partitioning events). We also defined two subsets of the LightDev, Stress, and Diff data sets using the first and last element of each times series respectively because the expression of genes at different points of a time series are potentially correlated. The number of genes with >5 partitioned conditions genes decreased in the subsets of LightDev (all = 334, first = 327, last = 325), Stress (all = 347, first = 265, last = 272), and Diff (all = 351, first = 277, last = 269) data sets. We excluded the Ctrl data set because it is composed of only four series, mean that no genes could pass the >5 partitioned condition cutoff.

The expected distribution of asymmetry values for the expression states of TF WGD-duplicates (under the assumption of independent of partitioning events) was determined by conducting a series of Bernoulli trails equal to the total number of partitioned states amongst TF-WGD duplicates. In each of these trials there was an equal probability that either the first or second duplicate receive the ancestral state. The results of these trials were then grouped according the exact per gene distribution of partitioned states in TF-WGD duplicates and an asymmetry value was calculated for each group. This procedure was repeated 1,000 times using an independent set of trials and subsequent groupings

For assessing *cis*-regulatory site asymmetry, only TF WGD-duplicates with ≥5 inferred ancestral *cis*-regulatory sites we considered (402 WGD-duplicate pairs total). Similar to expression state asymmetry, in each duplicate pair the ancestral and non-ancestral duplicates were defined according to the number of inherited ancestral sites. For each WGD-duplicate pair, the degree of asymmetry of *cis*-regulatory site among a TF pair was defined analogous to what

was done for expression. The expected distribution of asymmetry values for the *cis*-regulatory sites of TF WGD-duplicates was determined using the same procedure as for expression states.

Ordinary differential equation models of TF state evolution

The change in expression states from the ancestral expression quartile to either a higher or lower quartile in an extant TF was modeled as a system of ordinary differential equations such that:

$$\frac{d}{dt}\begin{pmatrix} O \\ + \\ - \end{pmatrix} = \begin{pmatrix} -(x+y) & w & z \\ x & -w & 0 \\ -y & 0 & -z \end{pmatrix}\begin{pmatrix} O \\ + \\ - \end{pmatrix}$$

Where *O*, +, and - are the frequency of TF WGD duplicate genes retaining the ancestral expression states, having a higher-than-ancestral expression level, and having a lower-than-ancestral expression level, respectively. The parameters *x, y, w*, and *z*, define the transition rates between these states. This system of equations was solved in Maxima (http://maxima.sourceforge.net/index.html) and best parameters for the observed distribution of duplicates pairs were determined using maximum likelihood estimates calculated with the bbmle package in R (https://cran.r-project.org/web/ packages/bbmle/index.html). Non-linear minimization was used to approximate an initial guess, although the actual initial parameters often needed to be adjusted to reach a convergent solution. The best fit parameters for this single duplicate expression state evolution model can be found in **Supplemental Table 4.8**.

The loss of ancestral expression states in a pair of duplicated TFs was modeled as a system of ordinary differential equations such that:

$$\frac{d}{dt}\begin{pmatrix} O \\ I \\ II \end{pmatrix} = \begin{pmatrix} -x & y & 0 \\ x & -(y+w) & z \\ 0 & w & -z \end{pmatrix}\begin{pmatrix} O \\ I \\ II \end{pmatrix}$$

Where *O, I,* and *II* are the frequency of TF WGD duplicate pairs where both, one, or neither duplicate retained the ancestral expression state. The parameters x, y, w, and z, define the

transition rates between these states. This system of equations was solved and the initial and best parameters were estimated in the same fashion as above. The best fit parameters for this pairwise expression state evolution model can be found in **Supplemental Table 4.9**. The same model was also applied to ancestral regulatory sites with *O, I,* and *II* representing the frequency of TF WGD duplicate pairs where both, one, or neither duplicate retained the ancestral regulatory site.

ACKNOWLEDGEMENTS

We thank Johnny Lloyd and Zing Tsung-Yeh Tsai for their advice regarding modeling duplicate retention and analyzing the importance of predictive features.

APPENDIX

Supplemental File 4.1: Linear equations used to predict odds of duplicate of retention in different WGD events across function groups

$$
\begin{aligned}
Odds(\alpha) = \quad & -0.370 * (Expression\ Mean, AtGenExpress) \\
& + 4.763 * 10^{-4} * (Expression\ Max, RNA) \\
& -140.6 * (Nucleotide\ Diversity, \pi) \\
& -1.073 * 10^{-3} * (Protein\ Length) \\
& -2.325 * (Expression\ MAD/Median, AtGenExpress) \\
& -0.190 * (Number\ of\ Domains) \\
& +4.786 \\
Odds(\beta) = \quad & -0.294 * (Expression\ Mean, AtGenExpress) \\
& + 6.745 * 10^{-4} * (Expression\ Max, RNA) \\
& -4.103 * (Expression\ Correlation, AtGenExpress) \\
& +2.686 * (Paralog\ D_n) \\
& -0.484 * (Number\ of\ Domains) \\
& +4.130 \\
Odds(\gamma) = \quad & -0.806 * (Expression\ Mean, AtGenExpress) \\
& + 4.329 * 10^{-4} * (Expression\ Max, RNA) \\
& -587.5 * (Nucleotide\ Diversity, \pi) \\
& -5.553 * (Expression\ Correlation, AtGenExpress) \\
& -0.133 * (Maximum\ Percent\ Identity) \\
& +5.282
\end{aligned}
$$

Supplemental File 4.2: Predicting WGD-duplicate retention status of individual genes using machine learning

Machine learning models for TFs, kinases, and all genes in the genome were generated to predict whether a gene in a particular group had a retained WGD paralog from either the α, β, and γ event, as small numbers of β and γ made the difficult to correctly classify on their own.. The machine learning was performed using the Random Forest algorithm implement in the R package randomForest (https://cran.r-project.org/web/packages/randomForest/index.html). We filtered the gene level feature set from a previous study (Lloyd et al. 2015) by removing those with missing values for ≥5% of genes. For the remaining features, missing values were imputed with the rfImpute algorithm in randomForest using 10 iterations of 500 trees. The final matrix of genes and features for TFs, kinases, and the whole genome can be found in Tables S7, S8, and S9, respectively. Using the imputed data set for each group of genes and for each WGD event, we ran the Random Forest algorithm 10 times with 500 trees (each time with 10 fold cross validation) and collected the resulting votes (retained or not) for constructing Receiver Operating Characteristic curves (ROCs). The importance of each individual feature was assessed using Mean Decrease in Accuracy (MDA), the average number of genes misclassified across multiple runs as a result of removing the feature in question. The statistical significance of the difference in values of a feature between WGD-duplicates and WGD-singletons was determined using Welch's t-test.

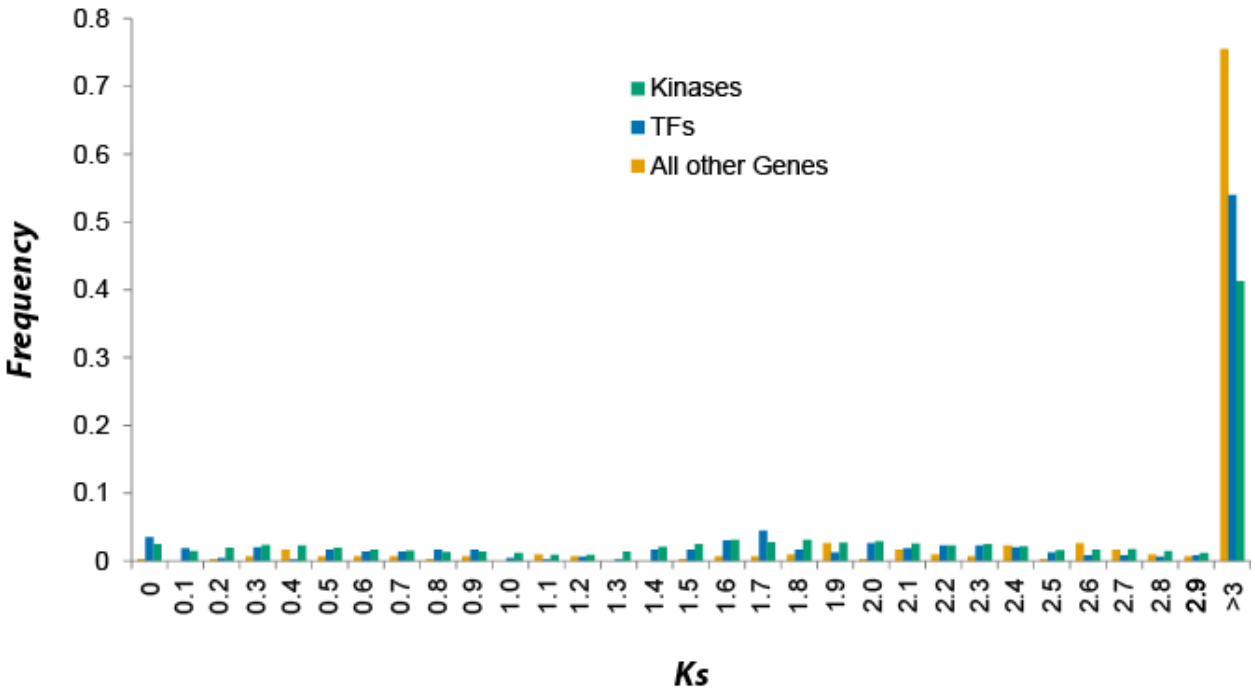
To evaluate the performance of our classifiers, we determined receiver operating characteristic curves (ROCs) for each model (Fig 1) and calculated the Area Under Curve (AUC-ROC), a metric that summarizes the ability of the classifier to recover true positive WGD-duplicate genes at different false positive rates. An AUC-ROC of 0.5 indicates that the classifier

is no better than randomly labeling genes as having a retained duplicate or not, while an AUC-ROC of 1.0 indicates that the classifier can make predictions without error. Among the classifiers, the one characterizing the full genome performed best (AUC-ROC = 0.86), followed closely by protein kinases (AUC-ROC = 0.82), while the classifier for TFs, while much better than random, did not perform as well (AUC-ROC = 0.74). To investigate the source of the difference, we determined the importance of each feature to the classifier by calculating the Mean Decrease in Accuracy (MDA) which is the average number of genes misclassified across multiple runs as a result of removing a feature (**Supplemental Table 4.10**). Given TFs are the least well predicted, we suspected the informative features for predicting retention in TFs would differ greatly from those for the genome at large and the protein kinases. Contrary to this expectation, the ranking of importance for TF WGD-duplicate prediction was more similar to the ranking of features for the whole genome prediction (Spearman's rank, $\rho = 0.86$) than the ranking of features protein kinases to the whole genome prediction (Spearman's rank, $\rho = 0.51$). This finding suggests that the feature value distributions of TF WGD-duplicate and WGD-singletons are more similar to the genome at large. Therefore, the reason that TF duplicate prediction model had lower performance was not simply because their feature values were substantially different from other duplicate genes. Instead, the features examined simply have lower importance in general for predicting TF retention (average MDA=11.3) than for other genes (average MDA=47.9), suggesting there are additional features important for TF retention that were not considered. For example, we might expect the number of DNA binding sites to be predictive of duplication status as an indication of the breadth of function of the TF which is related to the probability that a duplicate copy has been retained through subfunctionalization or gene balance.

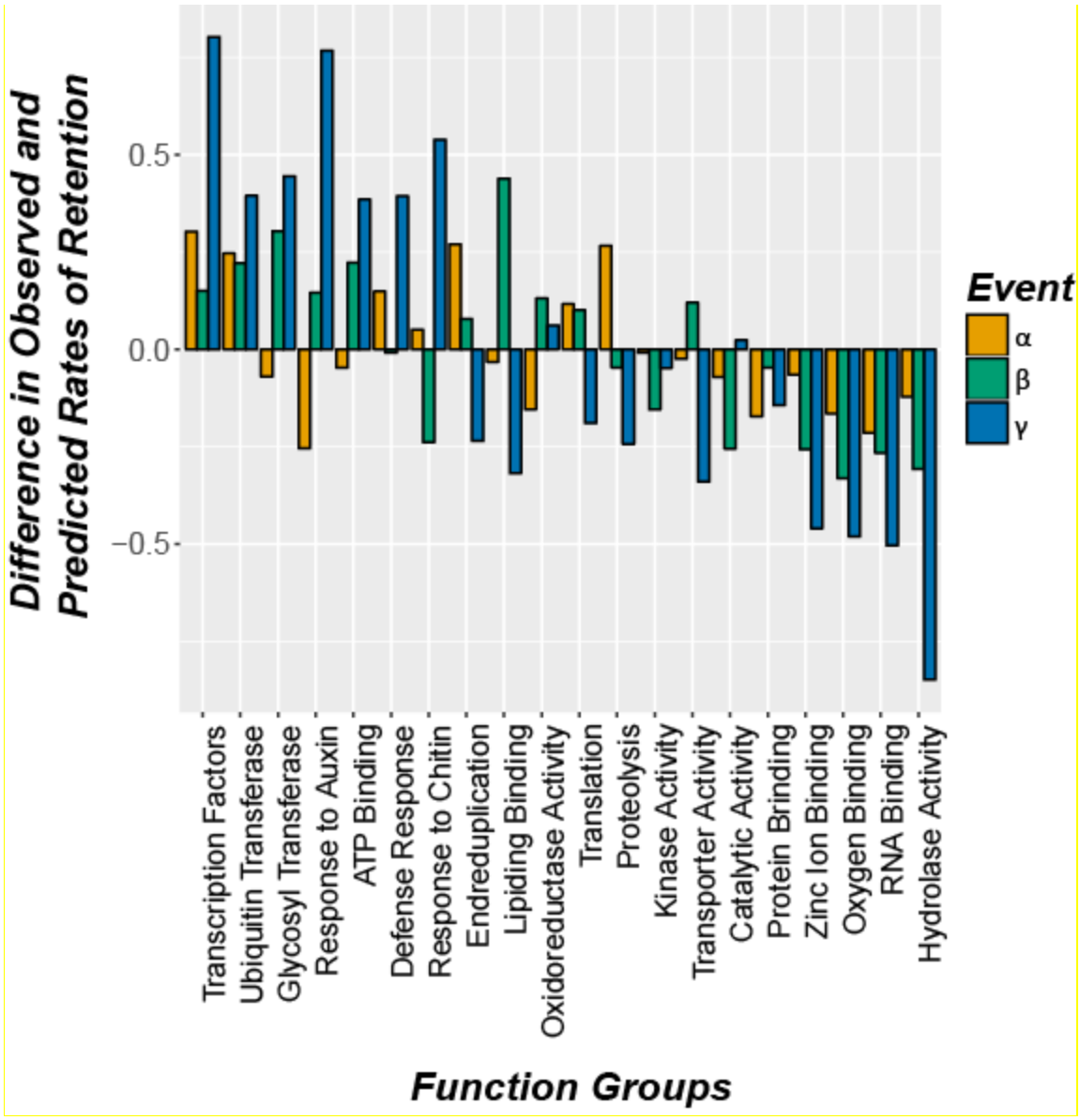
Furthermore, the most informative feature for classifying kinases and the whole genome, the percent identity to the best matching paralog in A. thaliana, was less important when applied to TFs (Table 3). Although the maximum percent identity of WGD-duplicates compared to WGD-singletons is significantly higher in full genome (pv = 1e-320), protein kinases (pv = 1.1e-36), and TFs (pv = 6.2e-12), the magnitude of the difference was greater for protein kinases (11.2%) and the whole genome (11.3%) than TFs (4.4%). This is due to WGD duplicate TFs having lower maximum percent identity (71.3%) than either kinases (75.2%, p=4.1e-24, t-test) or all genes (72.5%, p=5.9e-83 , t-test), while WGD-singletons TF had higher identity (66.9%) than kinases (64.0%, p=4.2E-35, t-test) and all genes (61.3%, p=1.9e-223, t-test). This observation may related to non-duplicate TF genes having apparent paralogs more often than non-duplicate genes do on average across the A. thaliana genome (Fig S2). The variance in the importance of maximum percent identity accounts for most of the performance difference across the classifiers as removing this feature yields similar results from all three (Fig S3). Similarly, inflating the difference in the percent identity of TF WGD-duplicates and WGD-singletons from 4.4% to 11.2% (the difference for protein kinases) would raise the predicted retention of TF from the γ WGD from 2.50 to 2.94, making up for more than half of the original error.

We would expect that other features used in our linear models would also be useful for classifying genes within function groups. However, the average importance rank of features found in more than one linear model was low (13.9 of 20), with the maximum expression value in RNA-seq being the worst feature in both the whole genome and TF classifiers. Of the four linear model features, mean expression in AtGenExpress had the highest rank in the whole genome (12th), TF (7th), and kinase classifiers (5th). However, the difference in mean expression between WGD-duplicates and WGD-singletons was not consistent: WGD duplicates

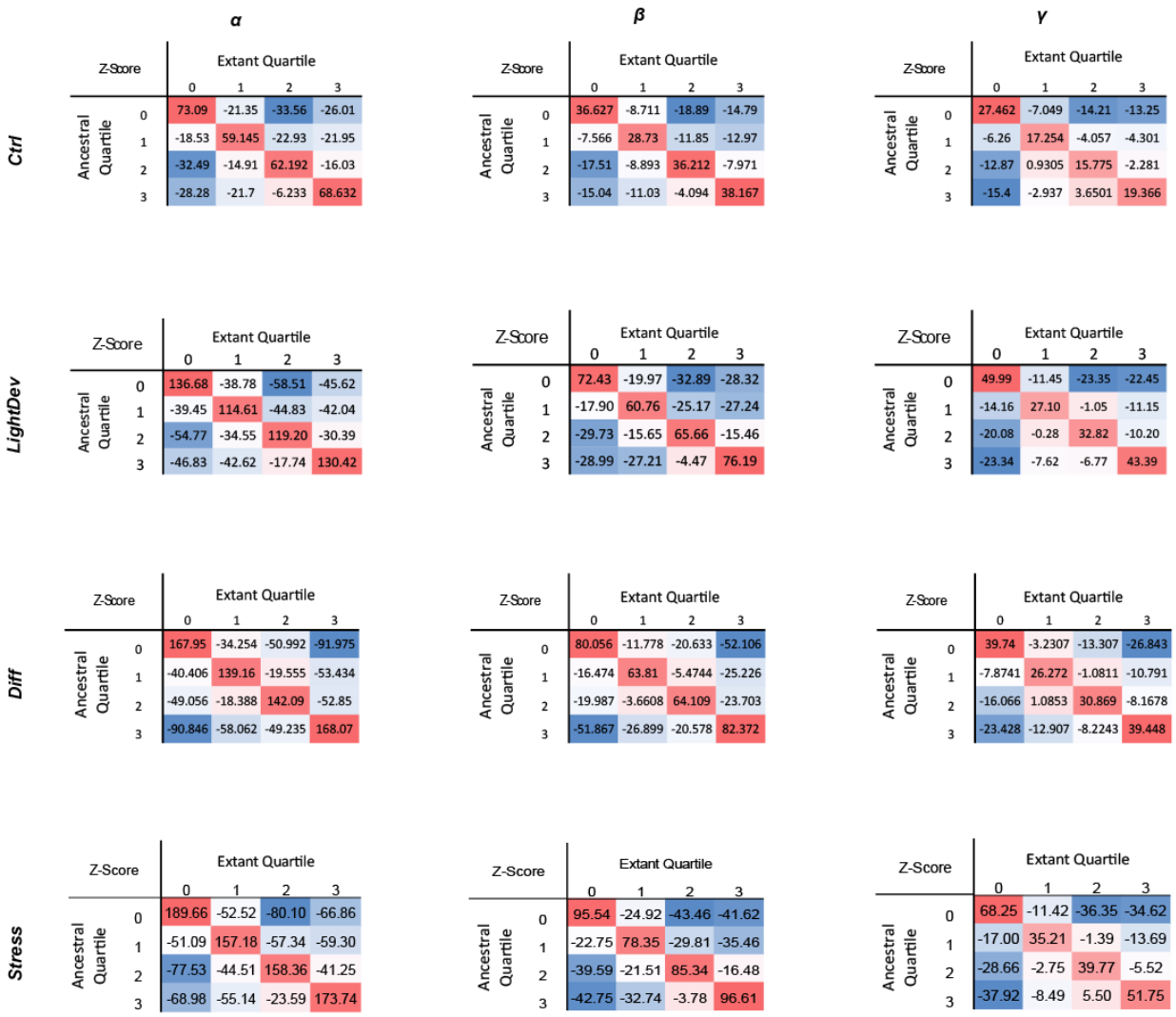
genes were more highly expressed across the whole genome (+0.32, p=4.0e-23), and TFs (+0.37, p=1.0e-4), but in protein kinases WGD-singletons were more highly expressed, though not at a significant level (-1.1, p=0.77). Hence, not only does relationship between gene features and retention depend on the gene function, but the relationship within individual function groups can be the opposite direction of the relationship across function groups. For example, the high retention of the TF function group is in part due to relatively low average expression in AtGenExpress, but within TFs, genes with higher average expression are more often WGD duplicates. This suggests that selection for duplicate retention is dependent not only on function and features, but their interaction as well, though the exact nature of these interactions is beyond the scope of this study.



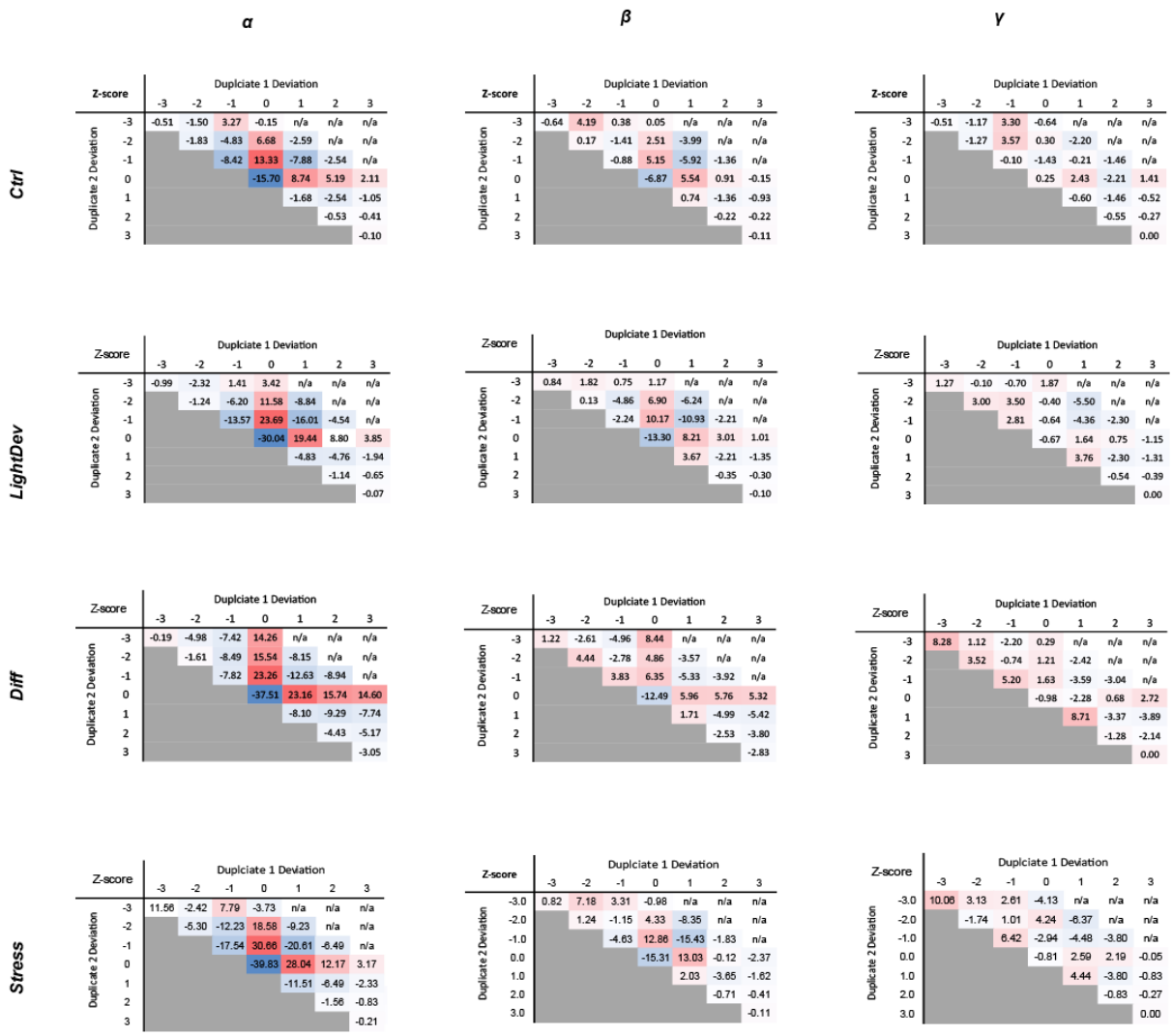
Supplemental Figure 4.1 Frequency distribution of synonymous substitution rate (Ks) between putative paralogs. Colors correspond to putative paralogs that are TFs (blue), kinases (green), or any other genes (orange). Known WGD and tandem duplicates are excluded from each group.



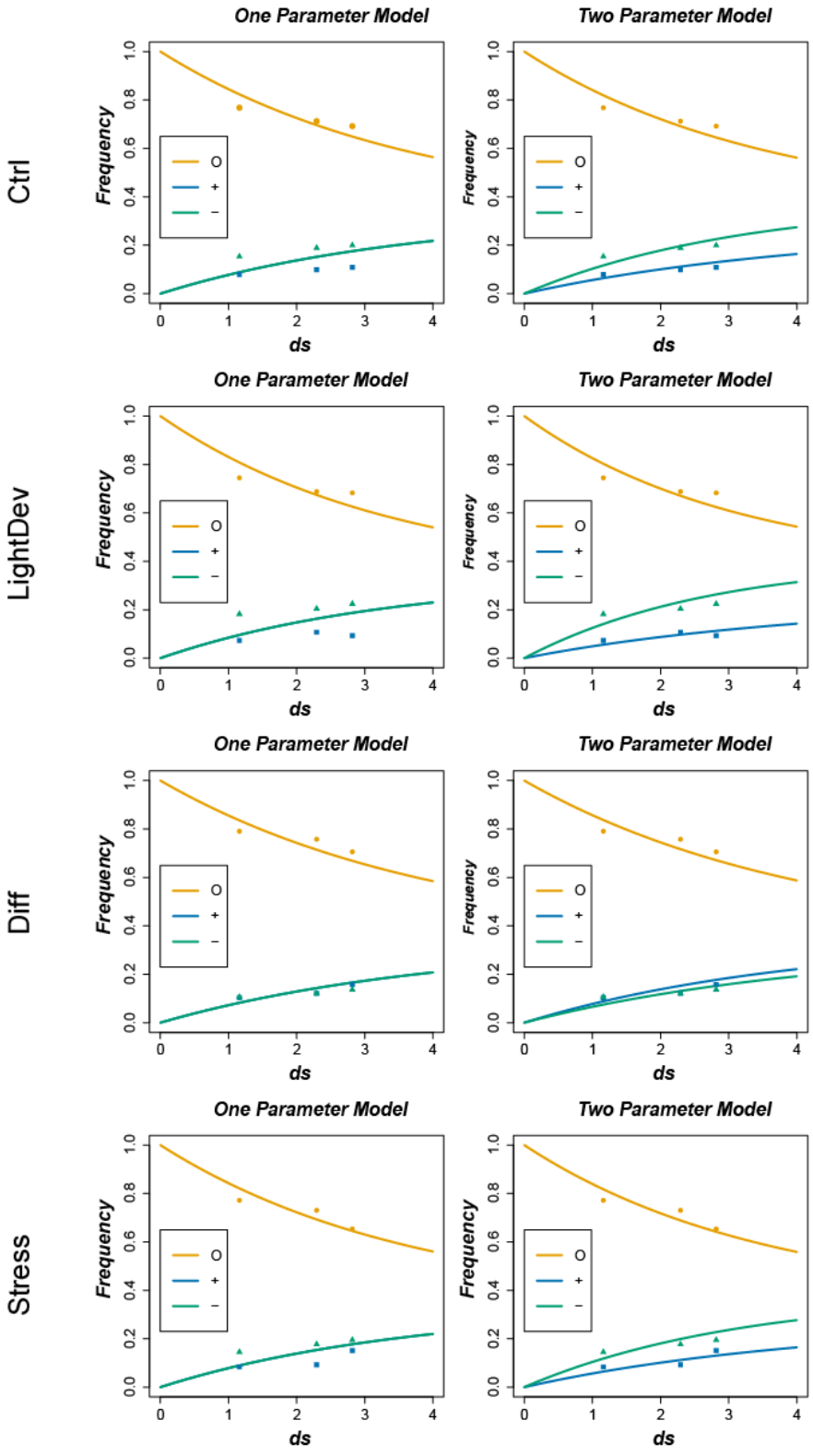
Supplemental Figure 4.2 Difference between the observed rates of duplicate retention and rates predicted by the linear models of duplicate retention. Different events are indicated by color (α = orange, β = green, γ = blue). Positive values indicate the observed rate is larger than the prediction while negative values indicate the observed rate is less than the prediction.



Supplemental Figure 4.3 **Difference in expression quartile of individual TF duplicates compared to their ancestral state.** Expression subsets (Ctrl, LightDev, Diff, and Stress) are indicated on the left and WGD event (α = left, β = middle, γ = right) along the top. Heatmaps show the z-scores of the observed frequency of each difference compared to the expected frequency. Color correlates with the magnitude of the z-score, with darker red values indicating counts further above random expectation and darker blue values indicating counts further below random expectation.



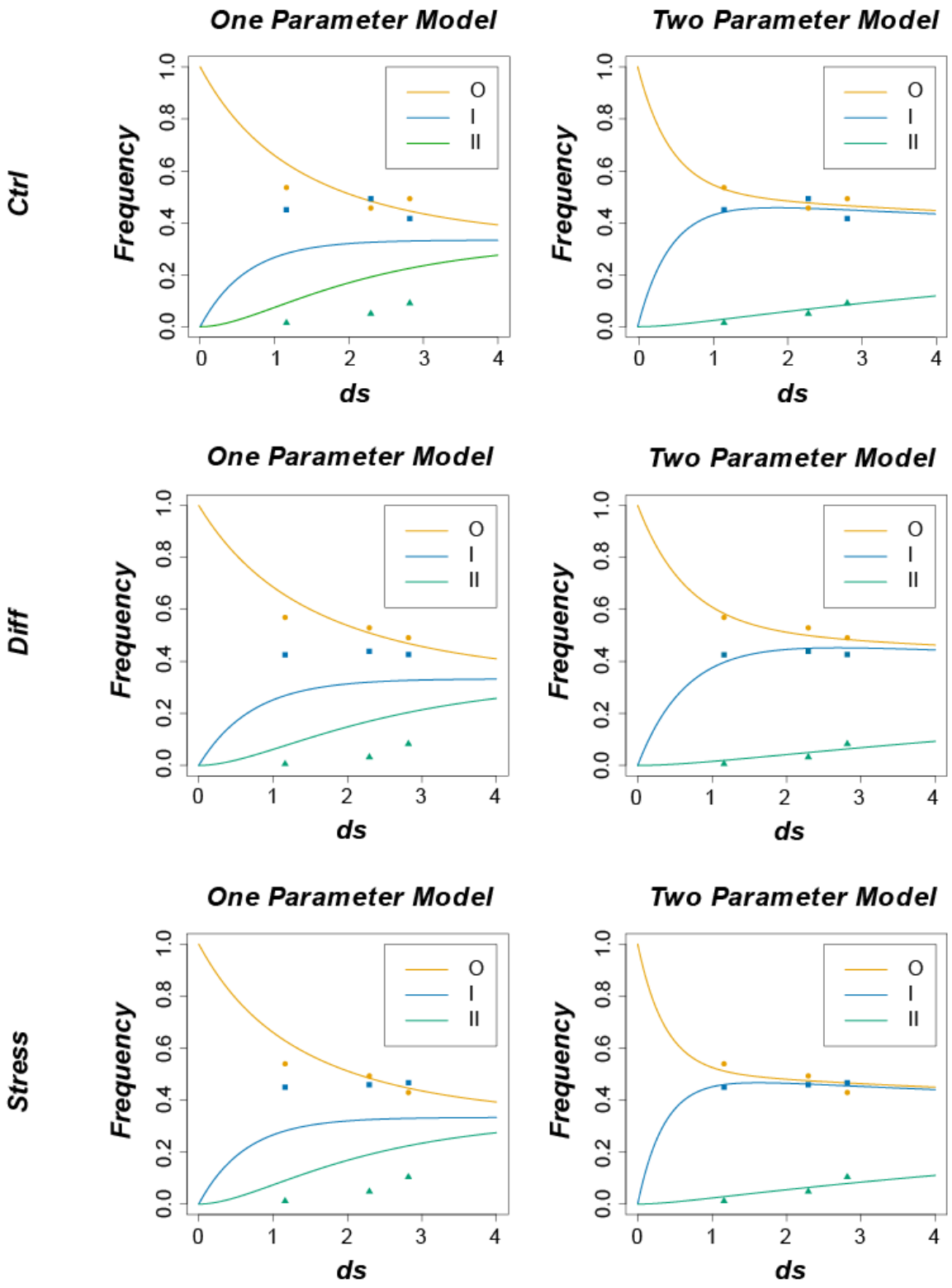
Supplemental Figure 4.4 Deviation of pairs of TF WGD-duplicates from their ancestral state. Deviation is defined as the difference value that each duplicate in a pair has from its ancestral state. Expression subsets (Ctrl, LightDev, Diff, and Stress) are indicated on the left and WGD event (α = left, β = middle, γ = right) along the top. Heatmaps show the z-scores of the observed frequency of the WGD-duplicate pair deviation compared to the expected frequency across all three duplicate events (α = top, β = middle, γ = bottom). Color correlates with the magnitude of the z-score, with darker red values indicating counts further above random expectation and darker blue values indicating counts further below random expectation.



Supplemental Figure 4.5 ODE models of the evolution of ancestral expression into either a higher or lower expression quartile. In this model, we consider the transition of a single WGD

Supplemental Figure 4.5 (cont'd)

duplicate from an ancestral expression state (O) to either a higher (+) or lower (-) expression state. Results for one (left column) and two (right column) parameter models show the change in time (x-axis) of the frequency (y-axis) of each state (O = orange, + = blue, - = green). Curves represent the continuous output of the model while symbols indicate the observed values on which the models were built (O = circle, + = square, - = triangle).



Supplemental Figure 4.6 ODE models of TF WGD-duplicate expression evolution relative to ancestral state for the Ctrl, Diff, and Stress expression subsets. In this model, we consider

Supplemental Figure 4.6 (cont'd)

the transition of the WGD-duplicate pair expression between three possible states relative to their ancestral state (O = both retained, I = one retained, II = neither retained). Results for one (left column) and two (right column) parameter models showing the change in time (x-axis) of the frequency (y-axis) of each WGD-duplicate-pair state (O = orange, I = blue, II = green). Curves represent the continuous output of the models while the symbols indicate the observed values on which the models were built (O = circle, I = square, II = triangle).

Supplemental Table 4.1 Data sets used in linear model of duplicate retention

Name	Description	Source	Use[1]
Gene Count	Number of genes in each Group	Internal	Kept
Paralog K_s	Synonymous substitution rate relative to the highest scoring BLAST hit *Arabidopsis thaliana*	Lloyd et al. (2015)	Dropped
Paralog K_a	Non-synonymous substitution rate relative to highest scoring BLAST hit in Arabidopsis thaliana (derived from paralog Ks and $\mathbf{K_a/K_s}$)	Lloyd et al. (2015)	Kept
Paralog K_a/K_s	$\mathbf{K_a/K_s}$ relative to the highest scoring BLAST hit Arabidopsis thaliana	Lloyd et al. (2015)	Dropped
K_a/K_s (A. lyrata)	Median $\mathbf{K_a/K_s}$ relative to genes in *Arabidopsis lyrata* in the same OrthoMCL Cluster	Lloyd et al. (2015)	Dropped
K_a/K_s (O. sativa)	Median $\mathbf{K_a/K_s}$ relative to genes in *Oryza sativa* in the same OrthoMCL Cluster	Lloyd et al. (2015)	Dropped
K_a/K_s (P. patens)	Median $\mathbf{K_a/K_s}$ relative to genes in *Physcomitrella patens*in the same OrthoMCL Cluster	Lloyd et al. (2015)	Dropped
K_a/K_s (P. trichocarpa)	Median $\mathbf{K_a/K_s}$ relative to genes in *Populus trichocarpa* in the same OrthoMCL Cluster	Lloyd et al. (2015)	Dropped
K_a/K_s (V. vinfera)	Median $\mathbf{K_a/K_s}$ relative to genes in *A. lyrata* in the same OrthoMCL Cluster	Lloyd et al. (2015)	Dropped
Expression Mean (AtGenExpress)	Mean of expression in AtGenExpress data	Internal	Kept
Expression Mean (Stress Data)	Mean of expression in StressTreatment subset of AtGenExpress Data	Internal	Dropped
Expression Mean (DevLight Data)	Mean of expression in Development and Light subset of AtGenExpress data	Internal	Dropped

Supplemental Table 4.1 (cont'd)

Expression Mean (Control Data)	Mean of expression in Control subset of AtGenExpress data	Internal	Dropped
Expression Mean (Diff Data)	Mean of expression in Stress Difference data set	Internal	Dropped
Expression Median (AtGenExpress)	Median of gene expression in AtGenExpress	Lloyd et al. (2015)	Dropped
Expression MAD/Median (AtGenExpress)	Median absolute deviation of expression over median expression using AtGenExpress	Lloyd et al. (2015)	Kept
Expression Breadth (AtGenExpress)	Number of AtGenExpress expression data sets with log2 intensity > 4	Lloyd et al. (2015)	Dropped
Expression Correlation (AtGenExpress)	Max of expression correlation with genes in the same OrthoMCL cluster using data from AtGenExpress	Lloyd et al. (2015)	Kept
Expression Correlation (AtGenExpress, Ks < 2)	Max of expression correlation with genes in the same OrthoMCL cluster that have Ks < 2 using data from AtGenExpress	Lloyd et al. (2015)	Dropped
Expression Module Size (AtGenExpress)	Size of co-expression module defined using K-means clustering of expression vectors from AtGeneExpress	Lloyd et al. (2015)	Dropped
Expression Breadth (RNASeq)	Number of expression in RNA-Seq Data set where the 95% confidence interval of FPKM does not include 0	Internal	Dropped
Expression Mean (RNASeq)	Mean of expression in RNA-Seq Data Set	Internal	Dropped

Supplemental Table 4.1 (cont'd)

Expression Median (RNASeq)	Median of expression in RNA-Seq Data Set	Internal	Dropped
Expression Maximum (RNASeq)	Maximum of expression in the RNA-Seq Data Set	Internal	Kept
Sequence Conservation (Viridiplantae)	Percent identity of BLAST hits with 34 plant species (described in Lloyd et al., 2015)	Lloyd et al. (2015)	Dropped
Sequence Conservation (Fungi)	Percent identity of BLAST hits with 8 fungal species (described in Lloyd et al., 2015)	Lloyd et al. (2015)	Dropped
Sequence Conservation (Metazoa)	Percent identity of BLAST hits with 8 metazoan species (described in Lloyd et al., 2015)	Lloyd et al. (2015)	Kept
Function Interactions (AraNet)	Number of functional interactions annotated in AraNet (http://www.functionalnet.org/aranet/)	Lloyd et al. (2015)	Dropped
Protein-Protein Interactions (AIMC)	Number of protein-protein interactions annotated in the *Arabidopsis* Interaction Network Map (http://interactome.dfci.harvard.edu/A_thaliana/)	Lloyd et al. (2015)	Dropped
Nucleotide Diversity (Pi)	Nucleotide Diversity (Pi) calculate between 80 *Arabidopsis* accessions	Lloyd et al. (2015)	Kept
Number of Protein Domains	Number of annotated protein domains	Lloyd et al. (2015)	Kept
Protein Length (in Amino Acids)	Length of gene's longest protein product	Lloyd et al. (2015)	Kept

Supplemental Table 4.1 (cont'd)

Maximum Percent Identity	Percent identity with the highest scoring BLAST hit in *Arabidopsis thaliana*	Lloyd et al. (2015)	Kept
Gene Family Size (OrthoMCL)	Number of genes in the same OrthoMCL Family	Lloyd et al. (2015)	Dropped

1. Indicates whether the feature was kept in the final linear model according to the filtering procedures described in Methods

Supplemental Table 4.2 Subsets of AtGenExpress used for ancestral expression inference

Data Set	Experimental Conditions	Total Samples	Number of Inferred States
Control	4	24	13600
Light and Development	34	91	47792
Stress Treatment	32	70	37391
Stress Differential	48	175	66602

Supplemental Table 4.3 Observed and expected frequency of duplicates TF pairs in a conserved, partitioned, and diverged state

Data Set	Duplicate State	α WGD		β WGD		γ WGD	
		Observed	Expected	Observed	Expected	Observed	Expected
Ctrl	Conserved	0.54	0.58	0.46	0.50	0.49	0.49
	Partitioned	0.45	0.37	0.49	0.42	0.42	0.42
	Diverged	0.01	0.05	0.05	0.08	0.09	0.08
LightDev	Conserved	0.53	0.58	0.47	0.51	0.49	0.50
	Partitioned	0.45	0.37	0.48	0.41	0.42	0.42
	Diverged	0.01	0.06	0.04	0.08	0.08	0.08
Diff	Conserved	0.57	0.62	0.53	0.57	0.49	0.51
	Partitioned	0.43	0.34	0.44	0.39	0.43	0.43
	Diverged	0.01	0.04	0.03	0.05	0.08	0.07
Stress	Conserved	0.54	0.59	0.48	0.52	0.46	0.47
	Partitioned	0.44	0.36	0.47	0.41	0.44	0.44
	Diverged	0.01	0.05	0.05	0.07	0.10	0.10

Supplemental Table 4.4 Experimental conditions used in each subset of AtGenExpress

Data Set	Conditions
Controls	Biotic control, Shoot control, Root control, Cell control
Light and Development	1st node, Carpels, Cauline leaves, Continuous blue light, Continuous darkness, Continuous far red light, Continuous red light, Continuous white light, Flower, Flower, Hypocotyl, Leaf 7, Leaf 7, Leaves, Mature Pollen, Mature Rosette, Mutant Rosette, Mutant Shoot, Pedicel, Petals, Red light pulse, Rif, Roots, Rosette, Seedling, Seed siliques, Senescing leaves, Shoot apex, Sepals, Stamens, Stem, UV-A-B light pulse, UV-A light pulse, Vegetative Rosette
Stress Treatment	avrRpm1, DC3000, Flg22, GST, GST-NPP1, H2O, HrcC, HrpZ, LPS, MgCl1, P, Psph, heat, MgCa, Root Cold, Root Drought, Root Genotoxic, Root Heat, Root Osmotic, Root Oxidative, Root Salt, Root UV-B, Root Wounding, Shoot Cold, Shoot Drought, Shoot Genotoxic, Shoot heat, Shoot Osmotic, Shoot Oxidative, Shoot Salt, Shoot UV-B, Shoot Wounding
Stress Differential	Root Cold 4C, Root Drought, Root Genotoxic, Root Heat, Root Osmotic, Root Salt, Root UV-B, Root Wounding, Root avrRpm1, Root DC3000, Root Flg22, Root GST-NPP1, Root HrcC, Root HrpZ, Root P. infestans, Root Psph, Root Cold 4C, Root Columnar, Root Cortex, Root Drought, Root Endo, Root Epi, Root Genotoxic, Root Heat, Root Osmotic, Root Oxidative, Root Protophl, Root Salt, Root Stele, Root UV-B, Root Wounding, Shoot avrRpms1, Shoot Cold4C, Shoot D3C300, Shoot Drought, Shoot Flg22, Shoot genotoxic, Shoot GST-Npp1, Shoot Heat, Shoot HrcC, Shoot HrpZ, Shoot Osmotic, Shoot P. infestans, Shoot Psph, Shoot Salt, Shoot UV-B, Shoot Wounding

Supplemental Table 4.5 RNA-seq data sets

Sequence Read Archive Identifiers
SRR1257404, SRR1257403, SRR1257402, SRR1257401, SRR1257392, SRR1257391, SRR1257390, SRR1257389, SRR976397, SRR976398, SRR976391, SRR976394, SRR929001, SRR929000, SRR921316, SRR921315, SRR921314, SRR921313, SRR921312, SRR921311, SRR671949, SRR671948, SRR671947, SRR671946, SRR653578, SRR653577, SRR653576, SRR653575, SRR653574, SRR653573, SRR653572, SRR653571, SRR653570, SRR653569, SRR653568, SRR653567, SRR653566, SRR653565, SRR653564, SRR653563, SRR653562, SRR653561, SRR653557, SRR653556, SRR653555, SRR649539, SRR649538, SRR649537, SRR634971, SRR634970, SRR634969, SRR584126, SRR584125, SRR584120, SRR584119, SRR520239, SRR520238, SRR520237, SRR515492, SRR515491, SRR515490, SRR515489, SRR479032, SRR477076, SRR477075, SRR452279, SRR452278, SRR452277, SRR452275, SRR452276, SRR452274, SRR445738, SRR445737, SRR441559, SRR441558, SRR402997, SRR402996, SRR402995, SRR402994, SRR391052, SRR391051, SRR070570, SRR070571, SRR069568, SRR069569, SRR069565, SRR069566, SRR069567, SRR069558, SRR069559, SRR069556, SRR069557, SRR974753, SRR974752, SRR974751, SRR974750, SRR652153, SRR652152, SRR652151, SRR652150, ERR274309, ERR274308, ERR274311, ERR274310, SRR1146545, SRR1055106, SRR1023821, SRR1020622, SRR1020621, SRR1005386, SRR1005385, SRR1005239, SRR1005238, SRR1001910, SRR1001909, SRR902025, SRR835483, SRR800754, SRR800753, SRR609268, SRR609267, SRR578948, SRR578947, SRR522012, SRR520363, SRR520364, SRR505746, SRR505745, SRR505744, SRR505743, SRR505137, SRR505135, SRR493043, SRR493039, SRR493036, SRR493033, SRR445736, SRR445735, SRR445214, SRR445215, SRR445216, SRR445217, SRR445218, SRR445209, SRR445210, SRR445211, SRR445212, SRR445213, SRR445204, SRR445205, SRR445206, SRR445207, SRR445208, SRR443169, SRR443165, SRR443164, SRR443163, SRR419186, SRR419182, SRR404277, SRR403907, SRR403903, SRR402370, SRR402371, SRR390314, SRR390312, SRR390313, SRR390311, SRR390310, SRR390308, SRR390309, SRR390306, SRR390307, SRR390305, SRR390303, SRR390304, SRR390302, SRR388670, SRR388669, SRR388668, SRR970149, SRR953401, SRR953400, SRR953399, SRR952321, SRR847505, SRR847506, SRR847503, SRR847504, SRR847501, SRR847502, SRR218098, SRR522916, SRR360147, SRR360152, SRR360153, SRR360154, SRR360205, SRR218099, SRR218100, SRR218092, SRR339951, SRR218101, SRR218102, SRR218096, SRR218097, SRR218089, SRR218090, SRR218085, SRR218086, SRR218087, SRR218088, SRR218094, SRR218095

Supplemental Table 4.6 Genes belonging to each GO-defined function group

Function Group	Genes
ATP Binding (GO:0005524)	AT1G14390, AT1G58050, AT1G01910, AT3G24240, AT2G24130, AT3G52570, AT5G26860, AT4G36180, AT4G17380, AT1G25320, AT5G49030, AT3G59760, AT5G44800, AT3G45300, AT2G23300, AT3G10350, AT3G57300, AT4G22730, AT4G01800, AT5G57450, AT4G36270, AT2G07040, AT5G63410, AT1G05120, AT4G02930, AT4G37870, AT4G36290, AT2G19860, AT3G10690, AT1G11100, AT2G26730, AT5G56030, AT2G27490, AT2G02090, AT5G58150, AT1G10850, AT2G45500, AT1G62750, AT5G17760, AT3G05780, AT1G32060, AT3G24340, AT1G74310, AT1G17750, AT3G28450, AT4G01020, AT5G51560, AT1G17290, AT2G42290, AT3G24660, AT4G23740, AT1G24290, AT2G28970, AT1G01220, AT5G38830, AT2G25790, AT3G46370, AT4G21800, AT1G70460, AT1G28440, AT2G21450, AT3G52200, AT5G21326, AT2G27600, AT1G62950, AT2G27170, AT3G28520, AT1G29900, AT5G65710, AT5G06820, AT3G53230, AT1G26190, AT1G49250, AT3G14350, AT3G20190, AT5G07660, AT3G49670, AT5G66760, AT2G13370, AT5G61030, AT1G74260, AT5G49770, AT3G56100, AT5G46330, AT3G19210, AT3G57760, AT1G75640, AT1G18130, AT2G15300, AT1G69990, AT4G00570, AT2G13800, AT4G24190, AT4G13850, AT3G27440, AT4G35740, AT5G03290, AT5G56040, AT4G34220, AT1G48650, AT3G54660, AT5G22370, AT5G40000, AT5G17730, AT3G62120, AT3G09660, AT5G63950, AT2G24230, AT4G25120, AT2G16250, AT4G28650, AT3G27190, AT2G26080, AT3G24495, AT4G10320, AT1G48480, AT3G13065, AT2G20420, AT1G74230, AT5G45840, AT1G74330, AT3G50940, AT1G58060, AT5G67520, AT5G14610, AT5G07810, AT5G61480, AT1G51980, AT4G26300, AT4G31180, AT2G32800, AT5G14470, AT2G02220, AT3G06480, AT3G06483, AT3G57640, AT2G30800, AT1G29750, AT5G04110, AT4G23240, AT4G18640, AT3G02130, AT5G10020, AT5G46280, AT3G18810, AT3G28610, AT5G16715, AT2G35120, AT3G51740, AT4G36280, AT5G65720, AT4G12790, AT5G56000, AT4G01900, AT3G20475, AT1G27880, AT1G63940, AT1G72180, AT4G23900, AT3G56370, AT2G26280, AT1G27190, AT1G73080, AT3G14840, AT5G61460, AT5G35390, AT1G78900, AT5G07440, AT4G37840, AT2G45280, AT2G36570, AT5G19310, AT5G52520, AT1G66830, AT1G53730, AT4G03390, AT2G25840, AT1G29870, AT2G27060, AT5G60730, AT1G72300, AT2G18470, AT5G48600, AT2G44350, AT3G16600, AT3G02880, AT1G09970, AT3G47110, AT5G44635, AT3G50930, AT2G01130, AT1G06840, AT5G20690, AT1G53420, AT5G24100, AT5G45800, AT1G14610, AT2G44980, AT3G50230, AT3G10270, AT3G13170, AT3G53590, AT1G14000, AT4G11010, AT3G22880, AT4G00960, AT5G44700, AT5G26742, AT1G64210, AT5G16050, AT2G41820, AT3G02660, AT4G23270, AT4G20270, AT1G07200, AT3G02065, AT1G51830, AT2G01950, AT1G79930, AT1G50410, AT4G04350, AT5G50920, AT1G72040, KATE, AT1G75820, AT2G01460, AT1G50480, AT5G22010, AT3G23890, AT5G01890, AT4G37250, AT5G63310, AT5G59660, AT1G55810, AT3G20010, AT3G54670, AT1G63680, AT1G51390, AT3G19700, AT5G58300, AT2G33840, AT3G56300, AT1G79620, AT3G06010, AT3G08680, AT3G28570, AT3G28580, AT2G01210, AT4G33760, AT3G42850, AT5G01950, AT4G35030, AT3G49240, AT1G73980, AT3G04600, AT3G47090, AT5G37450, AT1G72460, AT2G02780, AT4G12060, AT2G14050, AT2G14750, AT3G25840, AT5G14210, AT1G09620, AT3G54280, AT4G35520, AT3G42670, AT3G45450, AT5G49780, AT5G41180, AT3G47570, AT2G13560, AT4G31250, AT5G20480, AT3G28600, AT5G16590, AT4G37910, AT4G24830, AT1G65190, AT3G45770, AT5G65700, AT1G21650, AT3G06580, AT1G31420, AT1G43910, AT5G56010, AT1G78980, AT5G05130, AT3G23990, AT1G66530, AT2G18190, AT1G05910, AT2G18193, AT3G07770, AT1G48310, AT1G03030, AT2G31170, AT5G55200, AT5G04895, AT3G24320, AT2G18760, AT2G16440, AT3G12580, AT5G54590, AT5G51350, AT4G30250,

Supplemental Table 4.6 (cont'd)

	AT3G03770, AT1G74360, AT3G20040, AT2G16390, AT4G25370, AT5G47040, AT2G45590, AT4G20140, AT2G35920, AT3G28540, AT5G10880, AT2G04030, AT2G45340, AT5G17740, AT3G48870, AT1G72330, AT1G49270, AT1G63990, AT5G51070, AT3G24550, AT1G17410, AT1G68400, AT3G57830, AT2G25140, AT1G08600, AT4G29130, AT5G09590, AT3G47580, AT3G59410, AT1G08130, AT4G28490, AT3G13490, AT5G63710, AT3G48000, AT1G02670, AT1G67840, AT5G45780, AT2G33170, AT3G17840, AT1G07190, AT5G15920, AT1G33390, AT3G10700, AT3G03900, AT5G53320, AT2G46020, AT1G44900, AT1G50460, AT4G34200, AT3G12810, AT2G05710, AT5G05160, AT4G24280, AT5G54090, AT1G67510, AT3G05790, AT4G39940, AT5G48940, AT2G45490, AT1G22300, AT5G64580, AT4G16130, AT1G35710, AT2G39730, AT5G15450, AT4G14350, AT1G52290, AT2G07690, AT2G32850, AT1G35720, AT3G44740, AT5G06580, AT5G22750, AT1G56130, AT3G27730, AT2G33210, AT3G28040, AT2G46370, AT5G20420, AT4G02460, AT2G46620, AT5G19720, AT4G26510, AT5G04130, AT3G18524, AT1G66730, AT5G58720, AT1G28350, AT4G09320, AT1G61140, AT5G65690, AT5G40870, AT1G63430, AT5G43020, AT3G11710, AT3G29800, AT5G10370, AT4G29990, AT1G34110, AT2G26700, AT4G02060, AT4G39280, AT5G43530, AT3G01640, AT1G12460, AT3G55010, AT1G07650, AT1G50610, AT1G48030, AT5G08670, AT3G55400, AT5G63930, AT1G65070, AT5G50780, AT5G02820, AT1G47840, AT5G18170, AT3G42880, AT2G34560, AT4G39270, AT5G67200, AT1G45332, AT3G17240, AT5G38480, AT4G29380, AT2G31880, AT5G40010, AT3G58140, AT3G26560, AT5G53890, AT4G08920, AT1G60630, AT5G20040, AT5G17750, AT5G67280
catalytic activity (GO:0003824)	AT1G14290, AT1G64660, AT1G78050, AT1G50090, AT4G32790, AT5G20260, AT5G40270, AT3G55180, AT3G14790, AT4G12960, AT2G26000, AT2G31955, AT3G03990, AT2G04440, AT3G46440, AT2G38660, AT1G02270, AT4G02850, AT1G03210, AT5G37000, AT1G74290, AT5G19290, AT5G65280, AT5G62220, AT3G10572, AT4G10100, AT2G47630, AT4G14440, AT3G17365, AT5G44480, AT3G20650, AT2G35100, AT3G49680, AT3G55190, AT3G23820, AT3G62860, AT3G23940, AT1G08940, AT5G02970, AT4G12870, ARA2, AT3G26820, AT5G41250, AT4G34360, AT4G13360, AT4G29530, AT1G78500, AT1G29840, AT3G24030, AT2G29630, ARA1, AT5G03800, AT5G41650, AT2G43400, AT4G00620, AT4G22330, AT5G40290, AT4G14430, AT2G47760, AT5G57040, AT5G11130, AT1G12350, AT3G19710, AT4G12250, AT5G14180, AT2G25100, AT4G38800, AT3G10690, AT5G48010, AT1G53500, AT5G01260, AT1G13635, AT4G08170, AT3G03230, AT5G27410, AT1G63450, AT4G30440, AT2G20370, AT1G30620, AT3G14890, AT4G34700, AT3G05170, AT5G37530, AT1G76730, AT1G07645, AT1G27440, AT1G64185, AT2G22570, AT5G14980, AT4G33540, AT4G30540, AT2G01730, AT4G24340, AT3G57630, AT4G18270, AT1G10060, AT1G34380, AT5G38360, AT5G23220, AT2G32410, AT2G29040, AT2G45310, AT2G34770, AT1G01290, AT4G22756, AT5G57800, AT2G39725, AT4G33180, AT2G46370, AT1G13700, AT5G52810, AT1G25375, AT4G00110, AT1G22170, AT1G09935, AT1G74300, AT3G16700, AT4G12230, AT1G34270, AT5G64150, AT3G12290, AT5G19670, AT3G47560, AT1G68470, AT4G22580, AT5G11910, AT3G04390, AT4G16690, AT1G10070, AT4G25720, AT1G74680, AT4G13990, AT3G53520, AT2G25710, AT3G62830, AT3G13800, AT5G25820, AT2G28760, AT4G16210, AT3G26780, AT2G43980, AT5G16890, AT2G31990, AT5G04120, AT4G12890, AT1G65520, AT1G69640, AT5G26570, AT3G12260, AT3G24730, AT2G17280, AT2G47650, AT1G05350, AT4G22753, AT5G36150, AT2G20770, AT3G60510, AT2G39420, AT3G14990, AT1G48420, AT5G41120, AT2G23820, AT5G11610, AT1G37150, AT5G49570, AT5G62840, AT4G25434, AT3G45400, AT1G26160, AT3G42180, AT5G02080, AT5G08290, AT4G02860, AT4G15940, AT5G01580, AT5G22940, AT3G19820, AT1G74260, AT5G44930, AT2G32960, AT1G54570, AT4G38370,

Supplemental Table 4.6 (cont'd)

	AT2G34850, AT1G02190, AT5G22460, AT4G30530, AT3G07620, AT3G47590, AT1G67410, AT3G60910, AT4G30550, AT1G07080, AT5G25310, AT3G52050, AT1G17890, AT3G26840, AT5G11560, AT5G65780, AT1G78570, AT1G52160, AT1G76060, AT4G12900, AT3G03240, AT3G50520, AT5G61840, AT5G33290, AT1G52920, AT4G29120, AT4G38040, AT4G00600, AT2G32740, AT2G42160, AT5G59290, AT3G49360, AT5G63420, AT3G16260, AT4G20870, AT3G62810, AT1G34340, AT3G16190, AT2G37700, AT5G57850, AT4G15370, AT5G28840, AT3G58830, AT5G24400, AT5G41130, AT2G19550, AT5G11350, AT3G03650, AT5G23230, AT3G05190, AT1G12850, AT4G12110, AT1G08310, AT1G02000, AT4G20460, AT1G21480, AT5G61220, AT2G32750, AT5G22620, AT5G42600, AT1G50110, AT4G37500, AT5G16760
defense response (GO:0006952)	AT3G44480, AT5G06870, AT1G53350, AT5G44870, AT3G50950, AT5G51060, AT5G51700, AT1G12290, AT1G19610, AT1G61100, AT1G72840, AT4G04110, AT1G72920, AT1G64070, AT1G55010, AT4G02600, AT1G14410, AT5G25910, AT2G03300, AT3G46530, AT1G57830, AT2G39940, AT1G52040, AT1G61070, AT4G09360, AT3G05370, AT5G48620, AT4G16920, AT5G66910, AT1G69550, AT2G26380, AT2G15170, AT5G46470, AT5G17890, AT5G46270, AT5G56030, AT1G66100, AT5G63660, AT3G23180, AT2G32660, AT1G17615, AT1G63870, AT1G26700, AT4G11210, AT1G61310, AT3G26830, AT3G05650, AT5G23820, AT5G38340, AT3G51570, AT1G63750, AT1G57650, AT1G58400, AT4G36140, AT4G19510, AT2G38900, AT1G52660, AT1G09090, AT4G10780, AT5G18360, AT1G33560, AT4G08450, AT4G03550, AT1G72870, AT2G43510, AT5G43740, AT3G11820, AT4G02150, AT2G02100, AT5G45080, AT2G39200, AT4G13920, AT1G66090, AT1G12220, AT1G64160, AT2G33050, AT5G24780, AT5G41740, AT1G42560, AT3G24900, AT2G02740, AT2G23960, AT3G04220, AT3G13662, AT3G13660, AT1G79680, AT3G23120, AT2G32140, AT3G25010, AT5G64930, AT1G75830, AT5G44420, AT4G16890, AT1G59124, AT5G13160, AT5G18350, AT2G17430, AT4G16960, AT2G34930, AT1G71260, AT5G35450, AT1G12210, AT2G22330, AT1G61180, AT5G45260, AT5G45060, AT3G05660, AT2G03760, AT2G33670, AT3G46860, AT5G06860, AT5G44900, AT1G63360, AT4G09430, AT5G36930, AT1G58848, AT3G14460, AT3G46730, AT3G15010, AT1G72850, AT1G64060, AT1G72910, AT1G72520, AT3G49120, AT5G46760, AT3G51560, AT4G14370, AT3G45290, AT4G24230, AT3G53260, AT4G33300, AT3G09710, AT2G43520, AT4G23515, AT4G11170, AT1G66340, AT5G65970, AT3G05360, AT4G16930, AT5G17880, AT1G47890, AT2G02130, RPP22, AT4G13810, AT4G11190, AT5G46260, AT5G48780, AT2G02140, AT2G43910, AT1G17600, AT1G58390, AT5G04230, AT2G15010, AT3G23010, AT1G61300, AT1G63740, AT3G25020, AT5G44510, AT1G56520, AT2G15080, AT1G52900, AT5G38330, AT5G04720, AT1G58410, AT1G58807, AT4G19500, AT4G17880, AT5G42510, AT5G49040, AT4G04220, AT3G44630, AT1G60320, AT2G03030, AT3G50020, AT4G16940, AT2G14080, AT5G46520, AT3G20600, AT1G72260, AT1G17420, AT4G24250, AT2G15220, AT3G25510, AT1G65870, AT3G52450, AT1G19230, AT3G14470, AT1G50180, AT5G45090, AT2G33060, AT4G09420, AT4G38700, AT3G04210, AT3G23110, AT3G13650, AT5G46450, AT1G59780, AT1G65390, AT5G64905, AT4G26090, AT5G43580, AT5G17680, AT2G21100, AT5G07390, AT2G47730, AT5G05170, AT3G11080, AT5G45230, AT3G61220, AT5G40170, AT5G11250, AT1G72950, AT4G39950, AT3G15700, AT5G45070, AT3G56860, AT1G58602, AT1G56510, AT5G15130, AT2G35930, AT5G44910, AT1G58170, AT1G51480, AT5G40100, AT5G43730, AT5G41540, AT1G72860, AT1G72900, AT5G45200, AT4G39030, AT4G30070, AT5G48770, AT3G49110, AT4G16990, AT5G47910, AT5G40910, AT1G22900, AT4G19920, EDS9, AT4G19925, AT4G16900, AT1G63350,

Supplemental Table 4.6 (cont'd)

	AT1G15890, AT5G46490, AT5G58120, AT2G02120, AT4G11340, AT1G31540, AT4G11180, AT2G30860, AT3G48090, AT3G23240, AT5G47250, AT2G16870, AT1G27180, AT4G27190, AT1G63880, AT1G63730, AT1G45616, AT1G55210, AT5G53760, AT1G57670, AT4G19530, AT3G20820, AT3G44670, AT5G46510, AT1G72940, AT5G45240, AT1G09665, AT2G17050, AT4G26740, AT4G23570, AT3G11010, AT3G11480, AT1G57850, AT1G12280, AT3G45860, AT1G72930, AT2G41060, AT5G45220, AT3G16720, AT1G31580, AT1G10920, AT1G11000, AT2G32680, AT3G44400, AT1G47370, AT5G41550, AT4G19910, VET1, AT4G16950, AT5G66900, AT1G59620, AT2G23970, AT2G26010, AT5G43470, AT3G11340, AT2G21110, AT5G43570, AT1G57630, AT4G23310, AT5G47260, AT1G66980, AT5G18370, AT1G27170, AT2G14610, AT2G43710, AT1G59218, AT5G38350, AT5G36910, AT3G55230, AT5G22690, AT3G46710, AT5G51630, AT3G11840, AT5G45210, AT4G12010, AT1G61560, AT1G55020, AT5G49140, AT1G61190, AT2G17060, AT5G38850, AT1G73050, CIR3, AT1G72890, AT5G66890, AT1G33590, AT5G23400, AT4G23690, AT5G27060, AT4G23510, AT5G42500, AT2G37040, AT1G52030, AT3G07040, AT5G42650, AT1G71400, AT1G65850, AT5G44920, AT1G71390, AT4G23280, AT2G44110, AT1G12663, AT1G12660, AT5G55240, AT5G17970, AT5G13530, AT5G41750, AT4G27220, AT5G63020, AT4G36150, AT2G15130, AT2G26020, AT5G45000, AT5G47280, AT5G44430, AT5G15410, AT5G05400, AT1G56540, AT1G62630, AT4G16860, AT5G11270, AT4G19520, AT2G17480, AT5G45250, AT1G11310
DNA endoreduplication (GO:0042023)	AT3G21860, AT1G15570, AT2G27960, AT5G04470, AT2G19330, AT5G20570, AT3G20780, AT5G42190, AT1G20930, AT2G20140, AT4G26760, AT3G53970, AT3G08690, AT2G21550, AT1G49910, AT1G80370, AT1G70210, AT3G50070, AT4G05190, AT1G64520, AT1G47870, AT5G24630, AT2G22490, AT2G23430, AT1G69690, AT1G78770, AT1G77390, AT5G24330, AT3G13550, AT5G22220, AT1G75950, AT5G27620, AT4G34160, AT1G50490, AT3G11270, AT1G03780, AT3G48150, AT5G05560, AT5G11300, AT1G73690, AT1G66750, AT1G76540, AT5G48820, AT3G60010, AT5G41700, AT2G40550, AT5G57950, AT1G15660, AT2G16740, AT4G24820, AT4G28980, AT3G48750, AT3G12280, AT3G19150, AT5G10440, AT3G54180, AT1G64230, RFI, AT5G03415, AT1G20200, AT5G02470, AT3G42830, AT2G32710, AT3G24810, AT5G63610, AT3G25980, AT5G56150, AT3G20060, AT2G18290, AT2G42260, AT4G22910, AT2G39090, AT5G08550, AT2G03430, AT3G11520, AT3G15180, AT3G21850, AT3G48160, AT1G59540, AT5G11510, AT1G75990, AT1G47230, AT4G11920, AT5G64760, AT1G48380, AT5G13840, AT3G50630, AT5G09900, AT4G03270, AT5G51600, AT4G22970, AT2G27970, AT2G20000, AT5G65420, AT1G49620, AT4G29040, AT3G16320, AT4G14150, AT1G02970, AT1G06590, AT5G05780, AT1G29150, AT4G37630, AT2G36010, AT5G25380, AT1G50240, AT1G18040, AT3G60840, AT3G59550, AT4G38600, AT5G67100, AT3G19590
hydrolase activity hydrolyzing O-glycosyl compounds (GO:0004553)	AT1G11820, AT4G02290, AT4G27830, AT3G55260, AT4G19820, AT1G77780, AT3G57270, AT5G20870, AT3G60130, AT1G51470, AT1G75940, AT1G66270, AT5G58480, AT4G23560, AT5G16580, AT2G44480, AT3G09260, AT4G29360, AT1G64390, AT4G27820, AT4G19730, AT1G51490, AT1G75680, AT5G20940, AT5G09730, AT3G57260, AT5G20250, AT5G48375, AT2G05790, AT3G18080, AT3G62740, AT5G58090, AT3G47000, AT5G11920, AT3G60140, AT4G33810, AT3G13790, AT1G22880, AT4G22100, AT3G57520, AT1G33220, AT4G33830, AT4G33860, AT3G23770, AT5G20950, AT5G25980, AT2G20680, AT1G52400, AT3G30540, AT3G62750, AT3G47040, AT3G47010, AT2G44470, AT5G11720, AT2G19440, AT1G10050, AT4G19770, AT1G02310, AT1G55120, AT5G20390, AT4G08160, AT4G38650, AT3G19620, AT3G03640, AT2G14690, AT5G28510, AT3G43860, AT1G65610, AT2G01630, AT3G57240, AT3G55780, AT1G02640, AT3G04010, AT5G63840, AT3G55430, AT4G11050, AT1G12240, AT3G10900,

Supplemental Table 4.6 (cont'd)

	AT3G45940, AT2G44460, AT5G67460, AT1G70710, AT2G32990, AT1G66280, AT1G62660, AT4G19760, AT1G48930, AT2G39640, AT2G44490, AT5G42260, AT3G13784, AT1G13130, AT1G55740, AT1G71380, AT2G44550, AT4G18340, AT1G02850, AT4G17180, AT5G20340, AT5G49360, AT4G33840, AT3G23640, AT1G09010, AT5G64570, AT5G55180, AT3G61810, AT3G18070, AT4G26830, AT4G19750, AT1G30080, AT3G15800, AT4G21760, AT4G34480, AT4G19810, AT2G44540, AT3G26140, AT5G18220, AT4G39000, AT5G24540, AT1G78060, AT1G26560, AT1G60090, AT1G32860, AT3G07320, AT3G52600, AT3G46570, AT2G27500, AT4G19740, AT2G44450, AT5G36890, AT1G47600, AT4G31140, AT5G20330, AT1G05590, AT5G56590, AT4G01040, AT2G44570, AT5G16700, AT4G16260, AT1G61810, AT5G64790, AT4G33820, AT1G64760, AT3G54440, AT5G24550, AT5G49720, AT3G24330, AT4G14080, AT3G62710, AT4G09740, AT3G47050, AT2G25630, AT4G01970, AT2G36190, AT3G13560, AT2G16230, AT1G61820, AT2G44560, AT1G66250, AT1G23210, AT5G26000, AT1G02800, AT3G06510, AT5G20560, AT1G77790, AT4G38300, AT4G39010, AT5G01930, AT1G58370, AT5G54570, AT4G28320, AT3G10890, AT4G24260, AT5G44640, AT5G10560, AT1G68560, AT5G04885, AT5G40390, AT3G21370, AT1G19940, AT2G26600, AT4G19800, AT4G19720, AT5G24090, AT5G42100, AT5G66460, AT3G60120, AT5G42720, AT3G26130, AT2G32860, AT5G17500
kinase activity (GO:0016301)	AT4G40010, AT5G46080, AT1G51660, AT5G66850, AT2G25880, AT5G02290, AT1G01540, AT1G73500, AT5G27510, AT1G28390, AT3G17750, AT1G68400, AT1G67470, AT3G59350, AT2G02800, AT4G36180, AT4G10390, AT1G17160, AT1G25320, AT4G28860, AT1G09600, AT2G28940, AT5G15730, AT2G39180, AT3G48260, AT3G23340, AT1G29730, AT4G04500, AT2G23300, AT3G45430, AT4G28540, AT5G02070, AT3G57710, AT3G26940, AT3G63280, AT1G65250, AT4G23250, AT1G29230, AT4G23150, AT4G01330, AT1G12580, AT1G71530, AT5G08590, AT2G25760, AT2G14510, AT5G07620, AT5G10930, AT3G15220, AT2G21480, AT1G16440, AT5G59270, AT5G56580, AT1G03930, AT4G38230, AT1G51890, AT1G47890, AT5G65500, AT4G04570, AT2G24360, AT3G46330, AT2G41970, AT2G40560, AT3G53380, AT2G18530, AT2G40120, AT5G22050, AT4G33950, AT4G24480, AT2G43230, AT2G32660, AT4G23320, AT5G67080, AT3G12690, AT5G60310, AT2G34180, AT5G57035, AT4G38830, AT4G14780, AT3G13670, AT2G17220, AT3G23000, AT3G54180, AT1G22720, AT1G54610, AT5G10520, AT3G28450, AT3G05650, AT1G11350, AT1G09000, AT5G38990, AT2G42290, AT1G77720, AT1G70520, AT4G23740, AT4G13020, AT2G01460, AT5G04510, AT1G73460, AT2G28970, AT5G49470, AT3G50730, AT2G46700, AT1G16130, AT3G44200, AT2G25790, AT1G67000, AT1G56120, AT3G46370, AT3G44610, AT1G28440, AT1G67580, AT1G66920, AT3G19100, AT1G61550, AT5G66710, AT2G39110, AT1G62950, AT1G79670, AT5G60080, AT1G61380, AT4G39110, AT1G64630, AT2G17170, AT4G01370, AT1G33560, AT4G16970, AT5G61350, AT5G06820, AT5G66880, AT5G47850, AT1G26190, AT4G28880, AT3G01085, AT3G45670, AT3G20200, AT5G40540, AT1G67720, AT2G37050, AT1G53430, AT3G20190, AT1G06390, AT2G31390, AT4G28350, AT3G05050, AT1G61420, AT4G26610, AT4G08850, AT3G45790, AT3G01490, AT2G07020, AT3G10540, AT5G62230, AT1G33260, AT5G46330, AT4G13920, AT1G77280, AT1G11410, AT3G48750, AT5G18610, AT2G41930, AT5G35980, AT3G28040, AT2G33050, AT4G13190, AT1G52310, AT5G25910, AT1G75640, AT4G23160, AT2G15300, AT1G69990, AT1G14370, AT4G11480, AT4G02630, AT2G42960, AT1G51805, AT2G23950, AT1G79680, AT3G59790, AT4G33080, AT3G23120, AT2G33580, AT5G65530, AT2G45910, AT5G59650, AT3G24660, AT5G51830, AT3G20530, AT3G27440, AT4G34500, AT1G21230, AT5G56040, AT1G16120, AT1G18350, AT2G40500, AT5G41730, AT5G39440, AT2G30360, AT3G07980, AT5G13160, AT1G62400, AT5G54380, AT3G17410, AT1G70740, AT3G54030,

Supplemental Table 4.6 (cont'd)

AT5G60280, AT3G17840, AT1G20930, AT3G59480, AT2G16750, AT2G19230, AT2G37840, AT4G31110, AT1G61610, AT1G69270, AT3G27190, AT4G24100, AT1G24030, AT4G14580, AT4G23290, AT1G09440, AT3G17420, AT3G13065, AT1G78940, AT3G46930, AT4G04695, AT1G11330, AT1G61860, AT1G51810, AT1G73450, AT1G49350, AT5G01810, AT5G61550, AT1G53570, AT5G19450, AT3G04910, AT1G18670, AT1G61460, AT1G61500, AT1G19600, AT4G32830, AT2G04300, AT5G18190, AT1G16670, AT2G34290, AT4G08470, AT2G28930, AT1G61480, AT3G46160, AT2G46070, AT3G45440, AT3G22750, AT4G11530, AT2G05940, AT1G29750, AT1G32320, AT2G34650, AT3G46420, AT3G02130, AT1G51790, AT5G48380, AT5G07140, AT5G13290, AT4G21230, AT4G35310, AT2G14440, AT5G28290, AT3G05360, AT5G59260, AT5G67380, AT3G46400, AT1G03920, AT5G58520, AT3G51740, AT1G07870, AT3G21450, AT1G07150, AT4G18700, AT1G51800, AT5G01560, AT2G19190, AT3G59110, AT1G01450, AT1G18390, AT3G27560, AT3G05370, AT5G60550, AT5G55560, AT5G28680, AT2G41860, AT4G23230, AT3G13530, AT5G58140, AT1G23540, AT5G60890, AT1G72180, AT5G01060, AT4G11900, AT1G07570, AT4G04960, AT1G73670, AT1G18890, AT5G25110, AT5G58350, AT5G60320, AT2G43850, AT3G56370, AT1G27190, AT5G18910, AT4G04700, AT1G51870, AT2G23070, AT5G45430, AT5G24360, AT3G53640, AT2G20850, AT5G02800, AT5G11850, AT1G07880, AT4G04490, AT2G35620, AT2G38620, AT4G19110, AT1G55610, AT4G02410, AT3G50500, AT2G15080, AT3G12200, AT2G30980, AT1G21240, AT1G49730, AT5G12480, AT5G39000, AT3G25010, AT2G28960, AT2G28990, AT3G20830, AT3G50720, AT1G35670, AT1G80870, AT4G26890, AT1G66930, AT3G04810, AT3G55550, AT1G66830, AT1G03740, AT2G19470, AT1G25390, AT3G58640, AT4G21400, AT4G04220, AT2G38910, AT4G03390, AT5G11020, AT3G45860, AT1G69200, AT5G16900, AT4G18950, AT4G29050, AT3G05140, AT3G59730, AT2G38490, AT1G66750, AT5G35960, AT2G40270, AT3G55950, AT5G08160, AT1G05100, AT3G21630, AT2G39360, AT3G02880, AT3G11870, AT1G34210, AT1G09970, AT3G47110, AT1G17540, AT4G10730, AT3G45640, AT3G24540, AT5G28080, AT5G20690, AT1G53420, AT1G60800, AT1G61430, AT4G32660, AT3G53840, AT4G21940, AT3G02810, AT3G50230, AT3G59420, AT5G49760, AT5G51270, AT5G25930, AT3G45780, AT2G23450, AT4G11460, AT3G06230, AT5G59670, AT1G16260, AT1G06020, AT5G43910, AT1G12680, AT3G45410, AT2G36570, AT1G54960, AT5G60300, AT1G01140, AT2G40580, AT3G17850, AT2G41820, AT2G33060, AT4G11490, AT2G32510, AT1G18040, AT4G26540, AT1G65800, AT5G14720, AT1G10940, AT1G08720, AT1G48260, AT1G51830, AT4G13000, AT4G14340, AT2G43790, AT3G23110, AT5G62310, AT1G78290, AT3G22420, AT1G21590, AT3G09830, AT2G23030, AT1G64300, AT3G08870, AT3G57750, AT4G24740, AT2G30940, AT2G42550, AT1G63700, AT1G75820, AT4G31170, AT5G24080, AT5G04870, AT1G53050, AT5G67520, AT2G43700, AT3G11080, AT5G01890, AT4G37250, AT3G06620, AT4G23650, AT1G55810, AT1G16760, AT5G24430, AT5G40170, AT3G17510, AT5G58300, AT5G16000, AT1G66970, AT5G47070, AT4G18250, AT2G25440, AT3G13690, AT3G59830, AT1G70430, AT3G55450, AT4G26690, AT5G63940, AT1G79620, AT3G08680, AT3G58760, AT4G31100, AT3G51990, AT4G28706, AT1G61370, AT2G01210, AT3G53570, AT5G38560, AT4G35030, AT4G10010, AT1G49160, AT2G29220, AT2G17530, AT1G73980, AT3G51550, AT4G10260, AT1G16110, AT5G11400, AT1G60630, AT1G01560, AT5G37450, AT1G72460, AT5G18500, AT5G39030, AT3G20860, AT3G15890, AT3G62220, AT3G25490, AT3G09010, AT1G49180, AT5G58950, AT1G22870, AT5G19360, AT3G28690, AT3G25250, AT5G01820, AT5G41260, AT4G20450, AT1G29720, AT3G46340, AT3G53810, AT1G78530, AT4G08480, AT2G14750, AT1G68830, AT1G16160, AT3G58690, AT2G26830, AT5G01950, AT3G59740, AT1G72760, AT1G30640, AT5G07280, AT5G06740,

Supplemental Table 4.6 (cont'd)

AT3G14370, AT1G18160, AT3G46410, AT2G41140, AT5G20480, AT2G33020, AT1G15530, AT4G36450, AT1G76040, AT3G49060, AT4G32000, AT4G04710, AT5G63610, AT4G13820, AT1G11280, AT3G09780, AT3G53030, AT5G65700, AT2G30730, AT3G05660, AT1G74490, AT4G33430, AT4G18710, AT5G56460, AT4G29180, AT3G57740, AT1G31420, AT5G55830, AT1G01740, AT1G78980, AT5G41680, AT3G01300, AT2G41910, AT5G10290, AT3G44850, AT1G80640, AT1G48490, AT5G39420, AT4G23220, AT2G23080, AT1G66430, AT4G00710, AT5G65240, AT1G07560, AT3G04690, AT1G73660, AT1G51940, AT2G31010, AT5G35370, AT1G51860, AT1G30270, AT2G26290, AT2G37710, AT1G48480, AT5G58730, AT2G28250, AT3G08720, AT5G50180, AT4G11330, AT3G21340, AT3G52530, AT3G50530, AT5G44290, AT2G20300, AT4G35600, AT1G21250, AT1G53165, AT2G20470, AT4G35230, AT1G67890, AT2G47060, AT5G42120, AT5G51350, AT1G50700, AT1G69790, AT3G49370, AT5G26150, AT3G45390, AT4G27600, AT5G58540, AT2G18890, AT1G08590, AT5G11410, AT4G32300, AT3G46350, AT3G54090, AT5G43320, AT1G53700, AT4G23130, AT1G10620, AT5G40380, AT2G36350, AT5G22840, AT4G08800, AT1G61400, AT1G11300, AT5G42440, AT3G59700, AT2G07180, AT5G20050, AT3G16030, AT1G67520, AT5G38250, AT3G06640, AT2G29250, AT1G02970, AT5G61570, AT5G35380, AT1G04440, AT1G61590, AT4G32250, AT5G51770, AT4G14480, AT1G73080, AT5G60900, AT3G61960, AT2G44830, AT5G59680, AT4G23300, AT3G57830, AT1G26970, AT2G19410, AT2G31800, AT1G61440, AT5G37790, AT5G15080, AT5G45810, AT1G51170, AT2G19400, AT1G72540, AT5G50000, AT4G11890, AT1G16150, AT3G47580, AT5G41990, AT4G04510, AT4G28490, AT5G59700, AT4G02420, AT5G59660, AT5G65600, AT1G67840, AT3G57700, AT3G45420, AT2G32680, AT5G45780, AT3G45240, AT4G34440, AT2G24370, AT1G61950, AT1G65790, AT5G64960, AT5G61560, AT3G10660, AT4G23140, AT3G56760, AT3G01840, AT2G23200, AT4G00720, AT3G03900, AT3G08760, AT1G51820, AT1G48210, AT1G70530, AT5G63370, AT4G09570, AT1G59580, AT5G53320, AT5G25440, AT4G00330, AT5G44100, AT2G30740, AT4G23180, AT2G43690, AT5G38280, AT2G26980, AT1G54820, AT5G01540, AT1G74740, AT5G03320, AT3G18750, AT3G27580, AT1G52540, AT3G52890, AT3G20410, AT5G05160, AT1G70450, AT3G03940, AT1G53440, AT1G66460, AT5G63650, AT1G66980, AT1G49100, AT3G19300, AT5G01550, AT5G35580, AT1G51910, AT1G73690, AT5G09890, AT5G55090, AT4G27300, AT2G30040, AT1G51850, AT3G47090, AT1G61360, AT1G71830, AT2G28590, AT4G29810, AT1G10210, AT5G20930, AT3G61080, AT4G05200, AT5G12180, AT3G21220, AT1G76540, AT1G11340, AT2G17520, AT1G49580, AT4G22130, AT5G58380, AT5G23170, AT1G17910, AT4G14350, AT5G35410, AT5G60270, AT2G31500, AT1G57700, AT3G57530, AT1G34300, AT4G28980, AT5G39020, AT5G56890, AT3G57120, AT4G29450, AT1G17750, AT5G01920, AT4G21410, AT4G17660, AT5G57630, AT1G66910, AT5G58940, AT3G57770, AT2G11520, AT5G24010, AT1G69220, AT1G79640, AT1G54510, AT5G23580, AT3G59750, AT5G60090, AT1G05700, AT1G61390, AT4G03230, AT1G06700, AT2G48010, AT1G72710, AT1G67510, AT1G19390, AT4G02010, AT4G26510, AT2G25090, AT1G06730, AT5G27060, AT5G27790, AT5G56790, AT3G26700, AT4G25390, AT1G19090, AT2G05060, AT1G23700, AT3G46760, AT4G39940, AT3G53930, AT5G49660, AT5G40870, AT1G60940, AT5G66790, AT5G45820, AT5G35390, AT5G43020, AT4G23280, AT3G57720, AT5G10270, AT5G03730, AT4G04540, AT1G69910, AT3G56050, AT1G30570, AT5G07180, AT3G07070, AT2G23770, AT3G45330, AT2G18170, AT2G26700, AT5G18700, AT2G41920, AT1G79250, AT5G40030, AT1G64210, AT2G19210, AT4G04740, AT5G12000, AT3G46140, AT3G51850, AT1G51620, AT1G69730, AT4G25160, AT1G56720, AT1G07550, AT4G23210, AT5G03140, AT4G32710, AT5G48740, AT4G23190, AT3G23310, AT4G22940, AT1G70110, AT2G17290,

Supplemental Table 4.6 (cont'd)

	AT1G61490, AT2G39660, AT3G57730, AT1G11050, AT5G49780, AT3G63260, AT5G46570, AT1G70130, AT4G15530, AT1G45160, AT5G37850, AT1G29740, AT3G47570, AT1G03030, AT4G21380, AT1G50610, AT1G71410, AT4G31250, AT5G65710, AT3G08730, AT1G08650, AT4G24400, AT4G04720, AT4G30960, AT1G12460, AT2G17090, AT5G10530, AT3G42880, AT3G11010, AT5G55910, AT4G26100, AT5G57015, AT1G76370, AT1G55200, AT4G28670, AT5G01020**,** AT3G04530, AT5G67280, AT1G33770, AT5G03640, AT5G12090, AT5G58050, AT4G08500, AT4G39400, AT3G50310, AT5G16500, AT3G28890, AT2G25220, AT5G38240, AT3G24790, AT2G31880, AT5G07070, AT5G47750, AT2G35890, AT3G06030, AT5G53450, AT5G01850, AT1G06030, AT1G76360, AT1G51880, AT5G06940, AT4G26070, AT4G35500, AT5G16590, AT3G50000, AT5G03300, AT5G38260, AT3G46290, AT1G07650
lipid binding (GO:0008289)	AT5G07540, AT4G33355, AT2G15050, AT4G08670, AT1G43665, AT1G12100, AT1G62790, AT5G38170, AT4G27140, AT2G15325, AT1G55260, AT5G46900, AT2G27130, AT3G61050, AT3G18280, AT2G37870, AT5G07530, AT5G55460, AT1G32280, AT5G45560, AT3G22620, AT5G38160, AT5G55410, AT3G22580, AT2G48130, AT4G22630, AT4G27150, AT2G13820, AT4G12470, AT3G08770, AT1G73560, AT5G07520, AT5G56480, AT1G62500, AT5G48490, AT1G48750, AT3G57310, AT3G20270, AT4G30880, AT5G38195, AT4G22520, AT5G62080, AT3G22600, AT5G48485, AT1G62510, AT1G66850, AT4G22610, AT5G01870, AT5G07550, AT3G58550, AT1G04970, AT4G12490, AT5G38180, AT3G43720, AT5G13900, AT5G07230, AT2G10940, AT4G08530, AT3G52130, AT3G53980, AT3G22570, AT2G48140, AT3G51590, AT1G36150, AT4G12480, AT1G12090, AT4G12510, AT5G07510, AT4G22460, AT2G45180, AT5G46890, AT4G00165, AT4G12550, AT1G73780, AT5G54740, AT5G53470, AT1G73890, AT5G09370, AT4G22470, AT4G12520, AT3G22120, AT2G18370, AT5G07560, AT4G33550, AT2G44290, AT5G52160, AT4G19040, AT4G14815, AT4G27160, AT3G07450, AT1G73550, AT4G12530, AT4G12360, AT4G15160, AT4G22490, AT2G14846, AT5G05960, AT5G60690, AT5G64080, AT4G12500, AT4G27170, AT1G18280, AT5G55450
oxidoreductase activity (GO:001649)	AT3G03910, AT3G61580, AT3G05260, AT1G06350, AT3G61220, AT2G46210, AT1G14520, AT2G31360, AT5G04070, AT5G21482, AT3G03350, AT1G63380, AT2G29330, AT5G50600, AT1G20020, AT3G49620, AT5G49740, AT1G06100, AT1G49670, AT3G03100, AT5G06060, AT1G07440, AT4G10020, AT1G76150, AT3G20790, AT4G09670, AT1G72190, AT5G50690, AT1G06360, AT5G59540, AT3G03980, AT1G15140, AT5G49730, AT3G55290, AT4G20760, AT5G04900, AT3G49630, AT3G02280, AT3G06810, AT5G18210, AT2G29340, AT4G03140, AT1G07450, AT2G17845, AT5G07440, AT5G19200, AT1G51720, AT2G29290, AT3G01980, AT1G67730, AT4G24050, AT5G18170, AT1G03630, AT4G13250, AT3G08970, AT3G26760, AT3G50560, AT3G03330, AT5G50130, AT2G29350, AT2G47140, AT2G05990, AT2G24190, AT3G50210, AT2G22260, AT3G15850, AT1G75200, AT5G28310, AT1G52340, AT3G26770, AT5G02540, AT5G51030, AT5G50770, AT2G23096, AT4G23420, AT1G25460, AT2G29360, AT3G15870, AT5G10050, AT1G06090, AT1G15220, AT3G04000, AT1G12550, AT3G06060, AT3G21420, AT5G53090, AT2G29170, AT1G03990, AT3G60370, AT5G50590, AT4G23340, AT5G63290, AT4G26965, AT2G29370, AT1G62610, AT5G11330, AT5G60020, AT4G04930, AT4G17370, AT1G06080, AT3G46170, AT1G34200, AT2G47150, AT5G48440, AT4G05390, AT2G37540, AT5G56470, AT3G55310, AT5G61830, AT1G24470, AT3G47360, AT1G64590, AT3G51680, AT5G54190, AT1G79870, AT2G29300, AT1G57770, AT1G01800, AT2G29260, AT4G15093, AT1G68540, AT4G11410, AT5G64250, AT3G51840, AT5G67290, AT1G54870, AT1G10310, AT2G29150, AT2G47120, AT1G32480, AT4G05530, AT3G47350, AT4G23430, AT5G50700, AT2G07718, AT3G42960, AT1G58300, AT5G15940,

Supplemental Table 4.6 (cont'd)

	AT1G61720, AT1G06120, AT2G29310, AT2G38080, AT1G30510, AT4G09750, AT5G65205, AT2G30670, AT3G29250, AT5G53100, AT2G47130, AT5G66190, AT2G07727, AT3G29260, AT4G13180, AT2G29320, AT3G59710, AT4G16765, AT4G27440, AT3G56840, AT5G50160, AT4G27760, AT1G52810, AT5G60340
oxygen binding (GO:0019825)	AT4G31500, AT5G04660, AT3G48310, AT2G46660, AT4G39950, AT1G13110, AT3G20960, AT3G20140, AT2G30770, AT3G26180, AT4G15110, AT3G53280, AT2G34500, AT2G32440, AT1G55940, AT1G69500, AT4G37400, AT5G25130, AT4G37310, AT4G31950, AT1G74110, AT5G06905, AT2G12190, AT1G64940, AT3G14650, AT3G10570, AT2G29090, AT4G32170, AT5G51900, AT1G67110, AT1G33720, AT4G27710, AT1G28430, AT1G57750, AT3G20100, AT1G11680, AT3G26190, AT1G73340, AT3G56630, AT4G15350, AT4G19230, AT3G20090, AT1G12740, AT4G31940, AT4G37410, AT3G26290, AT1G64950, AT5G09970, AT5G25120, AT4G15360, AT2G44890, AT3G48520, AT4G37360, AT2G21910, AT2G14100, AT3G10560, AT3G14660, AT1G13090, AT1G17060, AT3G14620, AT4G12310, AT3G48300, AT2G30750, AT2G24180, AT3G53300, AT3G26280, AT5G23190, AT5G05690, AT3G20080, AT3G26150, AT1G01280, AT1G13080, AT1G01600, AT3G01900, AT1G58260, AT3G26200, AT5G44620, AT5G24900, AT2G23220, AT3G28740, AT3G26160, AT2G45550, AT5G36110, AT2G27010, AT5G24960, AT1G11610, AT5G14400, AT2G28860, AT5G25180, AT1G50520, AT5G10600, AT5G57220, AT4G15330, AT1G34540, AT2G27690, AT5G48000, AT2G46950, AT2G26170, AT5G08250, AT3G14680, AT3G26830, AT1G47620, AT1G75130, AT5G38450, AT5G47990, AT1G64930, AT2G27000, AT1G62580, AT5G24910, AT3G26270, AT5G67310, AT1G33730, AT1G74550, AT3G26170, AT3G26210, AT3G13730, AT1G01190, AT5G61320, AT2G42850, AT1G11600, AT4G39510, AT5G35715, AT3G19270, AT5G10610, AT2G28850, AT3G25180, AT1G16400, AT2G22330, AT5G04330, AT5G42580, AT5G38970, AT3G53130, AT5G45340, AT3G14690, AT3G48290, AT2G25160, AT1G64900, AT5G42650, AT1G13710, AT2G45570, AT3G53305, AT2G05180, AT3G14610, AT3G52970, AT3G26300, AT1G74540, AT5G04630, AT1G13140, AT4G13290, AT2G26710, AT1G65670, AT5G42590, AT2G45970, AT3G20110, AT1G24540, AT3G30290, AT5G07990, AT5G06900, AT3G26220, AT3G03470, AT4G13310, AT4G13770, AT5G52400, AT1G31800, AT4G37340, AT1G05160, AT4G37370, AT5G63450, AT1G66540, AT3G26310, AT4G12330, AT2G34490, AT1G13150, AT3G20120, AT4G15300, AT2G45580, AT3G20940, AT5G05260, AT2G46960, AT4G12320, AT4G15380, AT4G20240, AT2G16060, AT4G12300, AT4G22690, AT4G36380, AT3G44970, AT5G02900, AT4G00360, AT4G31970, AT3G48320, AT3G14630, AT3G61880, AT3G26125, AT1G19630, AT2G45510, AT4G39500, AT2G23180, AT1G65340, AT3G26320, AT1G13100, AT5G52320, AT3G20130, AT5G36130, AT1G50560, AT4G37330, AT3G20950, AT5G57260, AT5G25900, AT2G42250, AT3G14640, AT4G39480, AT3G53290, AT5G25140, AT4G22710, AT5G58860, AT3G44250, AT3G48280, AT1G63710, AT1G78490, AT4G37430, AT3G26330, AT5G36220, AT3G61040, AT3G30180, AT4G37320, AT5G24950, AT3G26230, AT1G79370, AT2G02580, AT3G48270, AT2G23190
protein binding (GO:0005515)	AT1G06190, AT3G21860, AT3G21865, AT1G76490, AT3G09770, AT2G43010, AT4G37150, AT1G23420, AT2G46280, AT1G16280, AT5G59710, AT5G33280, AT3G23820, AT4G02020, AT2G46260, AT5G13180, AT3G10670, AT3G01280, AT3G15150, AT3G56710, AT3G58040, AT1G50430, AT5G08130, AT5G65460, AT1G28520, AT3G05870, AT5G06950, AT5G45680, AT5G61960, AT5G51700, AT5G51120, AT5G16000, AT3G12690, AT3G62440, AT5G58440, AT5G27320, AT3G53120, AT4G35620, AT4G01026, AT5G23820, AT1G78080, AT5G37500, AT2G42830, AT5G04920, AT3G18730, AT2G32710, AT5G56860, AT1G15220, AT5G22330, AT5G15840, AT1G70700, AT5G57900, AT2G39090, AT1G10270, AT4G10920, AT2G25850, AT2G18840, AT2G38250, AT4G34000, AT3G28910,

Supplemental Table 4.6 (cont'd)

AT3G11820, AT2G34150, AT2G45980, AT5G49500, AT5G46330, AT4G00150, AT1G15910, AT4G35090, AT5G23310, AT2G41310, AT1G56650, AT1G25490, AT2G20080, AT3G11220, AT4G00570, AT4G20260, AT5G63880, AT1G02280, AT1G33410, AT2G44740, AT1G48270, AT2G17950, AT5G24520, AT5G43350, AT4G32650, AT4G19640, AT5G17690, AT3G24620, AT5G25780, AT3G50060, AT5G20570, AT4G26090, AT5G06140, AT2G47450, AT1G19120, AT5G35750, AT2G18710, AT2G27040, AT4G26160, AT5G40480, AT1G62360, AT1G28490, AT3G21870, AT5G64050, AT4G25320, AT5G41990, AT1G48500, AT5G40460, AT5G25760, AT3G14010, AT3G15540, AT4G27780, AT5G48820, AT5G43170, AT1G30135, AT3G06560, AT4G28270, AT2G46070, AT3G53260, AT5G13190, AT4G13250, AT1G71130, AT1G07270, AT3G60010, AT1G54320, AT3G56650, AT1G09530, AT5G16490, AT2G07560, AT1G31350, AT1G74890, AT5G44740, AT4G38130, AT1G30950, AT1G75410, AT1G69840, AT2G45770, AT4G28450, AT5G04930, AT4G30260, AT4G14700, AT1G24280, AT2G32720, AT3G48160, AT3G25500, AT4G14560, AT4G25530, AT5G03530, AT4G18290, AT1G74380, AT2G22810, AT1G74470, AT1G15100, AT5G37600, AT4G14147, AT4G20780, AT3G19290, AT2G14120, AT2G45190, AT5G46520, ATCG01130, AT1G73150, AT3G18980, AT3G45640, AT5G19400, AT5G67510, AT2G31380, AT2G20160, AT5G66570, AT5G08470, AT4G00360, AT3G45240, AT4G31730, AT5G14250, AT4G11880, AT3G16320, AT3G16857, AT1G60430, AT4G20870, AT1G17880, AT5G24400, AT1G01140, AT4G14960, AT1G13740, AT1G18040, AT3G19590, AT1G08720, AT2G40030, AT3G54170, AT4G11260, AT1G17080, AT5G56290, AT5G01410, AT2G19830, AT4G23650, AT5G01380, AT3G50070, AT2G22490, AT4G03190, AT5G24330, AT5G14920, AT1G02140, AT1G55310, AT5G64330, AT1G28480, AT2G29680, AT5G61010, AT1G77920, AT5G12900, AT4G17615, AT5G17020, AT4G28840, AT3G15000, AT1G06040, AT5G06150, AT5G21940, AT5G07280, AT5G48670, AT4G33270, AT5G55230, AT4G22920, AT2G24790, AT4G37770, AT5G61380, AT5G02470, AT3G06400, AT5G65800, AT5G20320, AT5G48250, AT4G26200, AT4G08980, AT3G10525, AT1G10470, AT4G38580, AT1G16330, AT2G26650, AT1G30490, AT1G29260, AT5G51100, AT3G17880, AT2G20890, AT3G06190, AT5G53160, AT5G43900, AT2G45740, AT3G14990, AT3G25810, AT4G35600, AT4G29910, AT4G14713, AT2G29570, AT3G61060, AT4G04780, AT2G38470, AT4G23810, AT5G16510, AT5G03520, AT4G39980, AT4G25540, AT2G44900, AT5G36250, AT1G13120, AT4G35040, AT5G37055, AT5G22220, AT4G14150, AT1G02970, AT4G14880, AT3G03450, AT5G51230, AT4G18130, AT2G36490, AT2G26150, AT3G02470, AT5G49450, AT2G17750, AT3G12360, AT1G15750, AT1G73830, AT5G10300, AT2G41680, AT3G63130, AT2G37630, AT5G42190, AT4G18620, AT5G42750, AT4G26840, AT4G24940, AT5G40930, AT1G19050, AT4G14550, AT5G63370, AT4G09570, AT1G47870, AT3G56980, AT1G17980, AT1G55520, AT3G04680, AT3G12810, AT4G19660, AT1G32230, AT5G05340, AT5G02110, AT4G25230, AT3G60250, AT5G27620, AT2G26300, AT5G51450, AT2G36960, AT2G40550, AT2G40000, AT5G05760, AT5G40280, AT4G09820, AT4G28980, AT4G27630, AT5G64070, AT3G54610, AT3G46710, AT3G27080, AT1G16890, AT5G55000, AT3G10730, AT4G26455, AT4G12720, AT1G53310, AT1G47220, AT3G57230, AT4G24540, AT4G17060, AT3G18524, AT2G37040, AT4G22910, AT3G21850, AT3G05420, AT2G30250, AT4G13520, AT4G13830, AT5G03730, AT1G64350, AT1G21700, AT3G50820, AT5G13530, AT1G30970, AT4G25160, AT1G20140, AT5G13840, AT1G51950, AT2G17290, AT4G21100, AT3G52750, AT3G52190, AT5G55910, AT2G33610, AT5G51200, AT1G24260, AT5G03280, AT1G50460, AT3G61070, AT3G19720, AT3G16050, AT2G31270, AT4G01090, AT5G16850, AT2G42880, AT5G61650, AT5G16690, AT4G25550, AT5G45250, AT2G36910, AT4G35050, AT1G45249, AT2G42400, AT5G22780, AT3G62410, AT2G31470, AT3G13300, AT3G45620,

Supplemental Table 4.6 (cont'd)

AT4G24972, AT1G32330, AT2G01120, AT5G41410, AT5G56580, AT4G11280, AT2G22540, AT2G38880, AT1G59610, AT5G51590, AT1G18080, AT5G25350, AT3G07780, AT5G47670, AT2G45820, AT2G27250, AT1G78600, AT5G56030, AT1G62170, AT5G41700, AT1G18400, AT1G22770, AT1G17200, ATMG00290, AT1G32060, AT5G10440, AT5G47010, AT4G30935, AT4G15630, AT5G41070, AT5G17790, AT1G17840, AT1G49620, AT1G69670, AT2G40010, AT1G31140, AT3G57130, AT3G54620, AT5G47700, AT5G35410, AT3G49850, AT2G46790, AT1G49950, AT1G49480, AT4G20940, AT1G31360, AT3G53230, AT4G02570, AT2G46225, AT2G43160, AT5G13300, AT1G14850, AT3G01770, AT3G15970, AT2G27100, AT3G50630, AT4G02150, AT5G44280, AT5G49910, AT4G27950, AT3G10540, AT5G48720, AT5G02100, AT3G57040, AT5G20850, AT3G20330, AT2G37180, AT5G02030, AT1G75080, AT5G03340, AT3G56400, AT3G52180, AT5G61850, AT1G68050, AT1G16610, AT4G37930, AT5G13160, AT5G52830, AT1G80840, AT4G10090, AT3G17510, AT4G25560, AT2G44920, AT1G20930, AT1G30210, AT3G53760, AT1G52740, AT1G10230, AT1G72450, AT1G04250, AT5G22570, AT1G01370, AT3G17609, AT5G08330, AT2G24120, AT3G52560, AT5G14750, AT3G48150, AT5G11300, AT1G48050, AT2G18020, AT4G32180, AT2G16770, AT3G22590, AT5G18580, AT4G24210, AT5G01840, AT1G19850, AT2G28060, AT4G09550, AT3G11730, AT5G63350, AT4G18700, AT5G24290, AT1G26670, AT3G22680, AT4G22200, AT5G60550, AT5G56280, AT3G51260, AT5G67250, AT3G03490, AT2G42810, AT4G08150, AT1G22985, AT1G66840, AT3G11130, AT2G17820, AT3G51860, AT1G03445, AT5G10450, AT4G17460, AT4G16110, AT3G23150, AT5G46860, AT5G09810, AT1G55610, AT1G01620, AT5G01590, AT1G35670, AT5G35620, AT4G23750, AT1G03840, AT1G64750, AT3G62980, AT5G13220, AT1G03190, AT2G31900, AT1G35580, AT5G39760, AT2G36010, AT4G27500, AT1G64990, AT2G46410, AT4G02560, AT5G65700, AT3G21630, AT4G24560, AT5G06850, AT5G44635, AT5G01820, AT3G01090, AT1G67710, AT4G16890, AT1G09570, AT5G05410, AT1G51510, AT5G13480, AT1G23900, AT4G04910, AT4G05420, AT1G50640, AT5G42970, AT1G65480, AT1G11400, AT1G31770, AT4G13180, AT4G25100, AT4G37630, AT3G46510, AT1G50240, AT4G13340, AT5G07070, AT1G64280, AT5G13790, AT5G62430, AT5G13820, AT1G58290, AT3G52770, AT1G65620, AT2G43790, AT5G45130, AT2G35110, AT3G01435, AT5G07090, AT2G26990, AT4G24740, AT4G29940, AT4G34210, AT3G10572, AT5G20900, AT5G65410, AT5G04990, AT4G25570, AT5G57050, AT3G28180, AT4G04770, AT3G09840, AT3G53570, AT4G15900, AT2G33560, AT5G09830, AT1G01560, AT5G27080, AT1G04240, AT1G01360, AT5G15290, AT3G25250, AT1G73590, AT5G66730, AT2G45790, AT2G38440, AT5G18260, AT4G11110, AT5G23430, AT5G62920, AT3G16650, AT5G11530, AT5G27150, AT1G04400, AT3G03000, AT5G63160, AT2G41140, AT1G61010, AT1G02170, AT3G24810, AT5G63610, AT2G26040, AT4G30820, AT1G02580, AT4G00850, AT4G18710, AT2G18160, AT2G20580, AT5G56540, AT4G32040, AT5G67260, AT4G27330, AT4G14830, AT3G02150, AT3G11630, AT3G23780, AT4G04720, AT3G15354, AT2G36100, AT3G22942, AT1G09415, AT1G09140, AT3G18165, AT5G14960, AT4G27420, AT1G77740, AT1G22190, AT4G23980, AT4G17870, AT5G05000, AT3G26090, AT3G03950, AT5G01900, AT4G08455, AT3G04740, AT4G12570, AT4G37130, AT3G60600, AT5G52250, AT4G12020, AT2G19560, AT1G09700, AT2G27050, AT5G01630, AT4G00355, AT5G65420, AT5G45860, AT4G02510, AT2G03160, AT4G02680, AT1G80080, AT4G37000, AT5G21274, AT1G16970, AT2G04240, AT4G32850, AT5G06200, AT5G55160, AT1G65700, AT3G21150, AT3G20310, AT2G33770, AT5G55300, AT4G02195, AT4G34110, AT1G50250, AT3G06530, AT3G01330, AT5G61600, AT1G65380, AT3G09920, AT3G20550, AT3G52890, AT1G20590, AT5G62390, AT1G71230, AT3G28860, AT4G39400, AT5G43830, AT1G73690, AT1G73000, AT5G61480,

Supplemental Table 4.6 (cont'd)

AT4G11850, AT3G21200, AT2G18915, AT1G01820, AT4G31160, AT1G08370, AT5G13860, AT3G62770, AT3G26060, AT3G52430, AT3G07040, AT5G63110, AT1G32900, AT5G10270, AT1G31440, AT5G46240, AT2G18170, AT3G03300, AT4G11920, AT2G41370, AT5G10470, AT1G79690, AT3G17860, AT3G46060, AT5G02820, AT4G00480, AT1G27630, AT1G75840, AT5G64920, AT2G19080, MPI1, AT4G27450, AT1G59660, AT3G47620, AT3G50310, AT3G54650, AT5G39340, AT5G47750, AT1G48760, AT1G76080, AT3G05590, AT2G03190, AT2G33310, AT4G26070, AT5G25380, AT3G58780, AT1G09770, AT1G63650, AT3G50000, AT3G54010, AT2G47430, AT4G12560, AT5G45870, AT3G20780, AT4G02500, AT2G03170, AT4G22950, AT4G33000, AT1G02450, AT2G22640, AT1G28250, AT5G60910, AT5G51660, AT1G74950, AT5G48930, AT3G08850, AT5G43060, AT3G15210, AT3G59760, AT5G02220, AT3G57090, AT2G06005, AT3G13550, AT2G39940, AT4G34470, AT2G13540, AT5G08590, AT4G37650, AT1G50710, AT5G13800, AT5G65720, AT3G01320, AT2G37560, AT1G69390, AT5G42480, AT1G30400, AT4G12780, AT4G30200, ATMG00090, AT4G33950, AT1G27930, AT1G07500, AT5G15160, AT1G75540, AT5G03220, AT3G12280, AT5G42080, AT5G45050, AT2G45640, AT1G24590, AT5G22290, AT5G60120, AT5G19280, AT5G62740, AT5G04510, AT1G04260, AT1G70660, AT3G05120, AT3G57860, AT3G02310, AT1G53720, AT1G14320, AT2G27600, AT2G45000, AT3G16830, AT2G18960, AT2G38310, AT4G01050, AT3G48100, AT3G28030, AT1G21410, AT3G19760, AT3G11540, AT5G66880, AT1G14920, AT3G09440, AT5G57360, AT3G26650, AT3G02000, AT3G49670, AT4G04020, AT2G34010, AT5G13120, AT5G63320, AT1G32640, AT5G24800, AT3G08530, AT2G41740, AT3G63010, AT1G05020, AT3G19210, AT2G36060, AT3G55000, AT3G55005, AT3G19180, AT3G02170, AT1G08180, AT3G24650, AT2G22430, AT1G12980, AT3G59790, AT1G77760, AT1G52890, AT5G27600, AT1G52380, AT5G60450, AT5G60100, AT3G51300, AT2G47510, AT5G40160, AT3G50670, AT5G50680, AT3G23050, AT4G16250, AT2G36270, AT5G56210, AT3G15730, AT1G17790, AT3G24495, AT2G40340, AT3G09260, AT5G06310, AT1G29680, AT3G57260, AT4G33010, AT5G05680, AT2G19430, AT2G36307, AT1G06110, AT5G50750, AT5G43070, AT1G09340, AT2G02560, AT2G05520, AT3G10340, AT2G36890, AT5G20810, AT4G34460, AT3G54840, AT5G38110, AT1G27320, AT2G34650, AT1G80350, AT4G37460, AT5G48380, AT5G13290, AT1G07370, AT2G30330, AT1G23860, AT1G52410, AT5G48160, AT5G45110, AT3G09900, AT3G21510, AT5G20920, AT3G28200, AT1G07880, AT3G20770, AT4G33945, AT5G03940, AT2G38620, AT5G58220, AT1G23260, AT2G42580, AT5G04230, AT3G52850, AT5G03540, AT1G10760, AT3G27920, AT1G53650, AT5G59880, AT4G24660, AT5G53490, AT3G14110, AT3G61570, AT2G38940, AT5G53470, AT3G06110, AT3G48680, AT5G49160, AT5G28770, AT1G47230, AT5G23880, AT5G23730, AT3G19770, AT3G11550, AT1G43850, AT2G05120, AT2G20000, AT4G23570, AT4G18040, AT4G36810, AT5G14620, AT5G23080, AT1G47128, AT3G18910, AT2G45450, AT3G45780, AT5G23260, AT3G22880, AT3G45140, AT4G29130, AT1G44800, AT4G32010, AT1G68640, AT3G05040, AT2G23350, AT5G67580, AT2G40750, AT5G66055, AT1G66390, AT1G49760, AT2G01950, AT3G07880, AT1G79830, AT2G18790, AT1G76310, AT5G63790, AT5G21010, AT1G75820, AT5G64900, AT1G69600, AT5G35840, AT3G43920, AT1G55490, AT5G63310, AT1G02860, AT1G12220, AT2G22240, AT2G02950, AT5G63860, AT4G19030, AT3G60890, AT1G09020, AT1G75950, AT3G56900, AT1G78870, AT3G43810, AT5G20240, AT1G44110, AT2G24765, AT4G17750, AT1G17745, AT3G14470, AT4G32551, AT5G43080, AT1G26840, AT1G31280, AT4G08480, AT4G29160, AT5G02200, AT4G02740, AT2G46340, AT5G07120, AT4G28560, AT5G03150, AT2G24490, AT2G01730, AT3G54850, AT2G13560, AT2G25700, AT3G60360, AT5G20480, AT1G23490, AT3G61140, AT1G30460, AT3G62090, AT1G80170, AT5G48170,

Supplemental Table 4.6 (cont'd)

AT4G35230, AT5G53480, AT1G21780, AT1G54040, AT3G12400, AT5G18620, AT2G21470, AT4G20380, AT1G73790, AT5G57160, AT3G26510, AT1G47750, AT1G10390, AT1G24150, AT2G30490, AT2G39770, AT5G12880, AT2G28380, AT5G16560, AT5G23720, AT3G48300, AT3G63140, AT5G14520, AT4G39710, AT3G62720, AT2G18300, AT4G36800, AT1G17760, AT1G73080, AT2G45280, AT2G41100, AT3G11410, AT5G11440, AT3G43440, AT2G46310, AT5G46210, AT4G35900, AT1G77140, AT3G62030, AT4G14870, AT3G29160, AT5G57740, AT3G48090, AT3G48360, AT4G23140, AT4G00650, AT4G20360, AT4G36480, AT1G32530, AT2G24840, AT5G05440, AT2G41430, AT3G23100, AT2G23430, AT5G11710, AT2G23380, AT3G25070, AT2G04660, AT5G39510, AT2G28800, AT4G24280, AT5G67480, AT5G64350, AT3G20000, AT1G55805, AT1G28420, AT4G19003, AT2G36250, AT3G51960, AT1G09030, AT4G26020, AT1G76540, AT1G02090, AT2G46970, AT4G03280, AT5G40740, AT3G56970, AT5G17620, AT4G02640, AT3G57530, AT1G63020, AT4G29510, AT5G20720, AT3G54220, AT3G15880, AT2G22670, AT3G25980, AT1G67580, AT2G26570, AT3G62800, AT1G50030, AT1G26830, AT4G29170, AT3G50360, AT1G53090, AT5G44200, AT1G27300, AT2G32400, AT3G15660, AT3G05380, AT1G06770, AT3G27010, AT5G13570, AT1G07130, AT4G27860, AT1G06950, AT4G13930, AT3G43300, AT1G16240, AT3G06590, AT3G53430, AT2G26700, AT4G02060, AT1G79250, AT2G46020, AT5G58590, AT1G29170, AT2G25000, AT5G50580, AT1G65290, AT2G31570, AT3G56370, AT1G30270, AT1G32310, AT4G32980, AT4G14310, AT4G37530, AT5G53290, AT4G09960, AT3G61600, AT5G03455, AT1G47580, AT5G58040, AT2G06530, AT3G29575, AT4G28640, AT1G25540, AT2G33460, AT3G53610, AT5G23860, AT2G39760, AT2G42010, AT2G17730, AT1G80680, AT1G73500, AT5G26980, AT5G15800, AT5G27100, AT4G18020, AT2G41110, AT2G16070, AT2G32950, AT1G04940, AT1G19180, AT2G22840, AT5G47100, AT1G01040, AT3G19220, AT4G15510, AT4G10760, AT5G27030, AT5G23040, AT5G41790, AT4G31580, AT2G40730, AT5G66030, ATCG00500, AT4G11140, AT3G48590, AT1G22070, AT4G36290, AT4G16420, AT4G33690, AT4G35800, AT4G19170, AT5G64960, AT3G55120, AT1G77080, AT3G25710, AT3G51970, AT5G63510, AT5G48570, AT5G52220, AT3G23000, AT3G54180, AT1G54610, AT2G37340, AT5G01370, AT5G47080, AT1G59750, AT2G01620, AT2G46700, AT3G04000, AT3G24440, AT4G32570, AT1G12390, AT3G03740, AT5G59430, AT1G71310, AT2G25490, AT1G35160, AT5G48400, AT1G69400, AT2G27370, AT3G24590, AT5G48990, AT5G55190, AT5G13490, AT3G25882, AT1G75010, AT3G15500, AT4G21560, AT5G20730, AT4G26610, AT2G31985, AT4G28910, AT5G55990, AT5G21170, AT5G20910, AT4G05000, AT3G18780, AT4G02070, AT1G08550, AT3G57290, AT3G13110, AT3G13445, AT1G70510, AT4G27920, AT4G35580, AT1G71860, AT1G34030, AT5G45010, AT2G01570, AT2G38170, AT4G29830, AT1G29010, AT1G75340, AT1G24310, AT3G14120, AT5G28640, AT2G30360, AT5G19000, AT5G11260, AT4G39800, AT5G15580, AT3G25230, AT4G30190, AT1G08830, AT4G35100, AT1G74500, AT5G13930, AT2G16850, AT4G38460, AT5G04240, AT1G04310, AT1G66740, AT5G11390, ATCG00190, AT5G16830, AT2G44950, AT5G08720, AT1G08780, AT2G30580, AT1G19350, AT3G02280, AT5G23670, AT5G46760, AT5G24020, AT3G63500, AT4G33510, AT1G72770, AT1G04550, AT5G27000, AT2G40890, AT1G66340, AT5G40810, AT5G10350, AT3G23670, AT5G24110, AT4G01900, AT5G09250, AT5G60410, AT3G33520, AT1G56330, AT5G24590, AT3G61190, AT2G26280, AT2G15400, AT1G05200, AT2G40380, AT3G50500, AT2G01980, AT2G40470, AT3G07560, AT1G77180, AT3G51630, AT5G40030, AT1G17730, AT1G02340, AT2G33270, AT4G26750, AT2G26350, AT5G06960, AT4G29350, AT1G76920, AT4G12620, AT4G26570, AT2G01760, AT3G57350, AT2G46600, AT3G46580, AT3G20600, AT3G47500, AT4G37520, AT1G02840, AT1G09270, AT5G65930, AT4G19990,

Supplemental Table 4.6 (cont'd)

	AT5G20010, AT4G14180, AT5G08450, AT1G69120, AT2G34180, AT5G26860, AT5G66750, ATCG00020, AT3G49700, AT1G47260, AT5G66280, AT5G07300, AT1G05460, AT2G30470, AT1G01160, AT1G69690, AT5G57450, AT1G78770, AT4G01370, AT5G18410, AT5G04140, AT1G22920, AT2G04550, AT3G43700, AT1G65030, AT3G02230, AT1G71830, AT2G45650, AT4G08920, AT1G01910, AT5G09800, AT3G62420, AT1G10690, AT4G14220, AT5G19330, AT3G19150, AT1G74310, AT3G57870, AT5G09790, AT3G59220, AT5G03415, AT3G13920, AT1G46408, AT5G20930, AT2G42260, AT2G24540, AT3G20740, AT1G70310, AT1G10970, AT5G58230, AT1G04450, AT3G06720, AT5G12390, AT1G71800, AT2G05380, AT2G45660, AT3G21430, AT2G04890, AT3G14080, AT1G21250, AT5G02500, AT3G21640, AT5G22880, AT3G47430, AT2G26760, AT4G31800, AT1G10940, AT4G02440, AT3G10650, AT5G04470, AT2G27960, AT5G16620, AT1G15570, AT5G58550, AT2G30770, AT1G21750, AT5G44550, AT5G04870, AT1G65650, AT1G70210, AT4G14110, AT5G57110, AT5G62000, AT5G41315, AT5G53120, AT2G21240, AT5G54490, AT5G04900, AT3G17590, AT3G12250, AT5G22640, AT4G09010, AT1G66750, AT4G04740, AT3G29350, AT5G42520, AT2G45140, AT3G16770, AT1G05560, AT2G19110, AT3G48750, AT1G73360, AT1G01510, AT1G21970, AT1G61790, AT1G53510, AT2G40940, AT1G22275, AT2G41620, AT5G10030, AT5G17560, AT4G37490, AT4G33430, AT3G63400, AT2G18040, AT5G09260, AT3G16420, AT2G40330, AT4G03080, AT5G40330, AT4G30960, AT5G25890, AT1G76260, AT5G44180, AT5G20020, AT5G24270, AT5G46790, AT5G35790, AT1G48380, AT1G32500, AT1G56250, AT1G49720, AT3G05280, AT4G03270, AT4G17710, AT2G44610, AT5G52550, AT2G36350, AT1G48410, AT2G46830, AT4G33650, AT1G80490, AT4G27160, AT1G17380, AT3G04730, AT1G62300, AT3G46590, AT1G12110, AT1G01640, AT2G36990, AT1G64860, AT4G32910, AT3G08900, AT5G55280, AT5G01640, AT3G57560, AT5G65210, AT5G54640, AT5G64630, AT5G22770, AT3G13460, AT1G03000, AT5G62810, AT5G02420, AT5G08120, AT2G27970, AT5G16320, AT4G13870, AT4G37940, AT5G59380, AT5G44560, AT2G38280, AT3G18690, AT4G39890, AT2G29100, AT1G74740, AT1G20610, AT4G29810, AT4G34160, AT3G06380, AT5G13060, AT5G15210, AT3G61630, AT3G44530, AT5G02490, AT1G30330, AT4G04885, AT2G05210, AT2G29210, AT1G79040, AT1G20780, AT4G14720, AT3G55840, AT1G79280, AT2G31500, AT5G14270, AT5G57380, AT5G15850, AT3G59380, AT1G10060, AT5G49060, AT3G54820, AT1G24510, AT1G02980, AT5G18400, AT4G36930, AT2G18290, AT4G23450, AT4G00180, AT5G67570, AT2G20180, AT4G31710, AT3G11520, AT5G14070, AT2G28160, AT5G50950, AT1G14740, AT3G23380, AT4G15410, AT1G26110, AT2G40670, AT1G80670, AT1G43700, AT4G15090, AT1G32400, AT3G51920, AT3G08730, AT2G01830, AT3G23030, AT4G30840, AT5G47120, AT1G09070, AT2G35720, AT5G52120, AT2G32980, AT1G56260, AT4G08500, AT4G26080, AT1G12360, AT3G54710, AT2G36160, AT5G25220, AT4G17490, AT3G50410, AT4G16144, AT5G39660, AT5G61210, AT5G60340
proteolysis (GO:0006508)	AT4G31500, AT5G04660, AT3G48310, AT2G46660, AT4G39950, AT1G13110, AT3G20960, AT3G20140, AT2G30770, AT3G26180, AT4G15110, AT3G53280, AT2G34500, AT2G32440, AT1G55940, AT1G69500, AT4G37400, AT5G25130, AT4G37310, AT4G31950, AT1G74110, AT5G06905, AT2G12190, AT1G64940, AT3G14650, AT3G10570, AT2G29090, AT4G32170, AT5G51900, AT1G67110, AT1G33720, AT4G27710, AT1G28430, AT1G57750, AT3G20100, AT1G11680, AT3G26190, AT1G73340, AT3G56630, AT4G15350, AT4G19230, AT3G20090, AT1G12740, AT4G31940, AT4G37410, AT3G26290, AT1G64950, AT5G09970, AT5G25120, AT4G15360, AT2G44890, AT3G48520, AT4G37360, AT2G21910, AT2G14100, AT3G10560, AT3G14660, AT1G13090, AT1G17060, AT3G14620, AT4G12310, AT3G48300, AT2G30750, AT2G24180, AT3G53300, AT3G26280,

Supplemental Table 4.6 (cont'd)

	AT5G23190, AT5G05690, AT3G20080, AT3G26150, AT1G01280, AT1G13080, AT1G01600, AT3G01900, AT1G58260, AT3G26200, AT5G44620, AT5G24900, AT2G23220, AT3G28740, AT3G26160, AT2G45550, AT5G36110, AT2G27010, AT5G24960, AT1G11610, AT5G14400, AT2G28860, AT5G25180, AT1G50520, AT5G10600, AT5G57220, AT4G15330, AT1G34540, AT2G27690, AT5G48000, AT2G46950, AT2G26170, AT5G08250, AT3G14680, AT3G26830, AT1G47620, AT1G75130, AT5G38450, AT5G47990, AT1G64930, AT2G27000, AT1G62580, AT5G24910, AT3G26270, AT5G67310, AT1G33730, AT1G74550, AT3G26170, AT3G26210, AT3G13730, AT1G01190, AT5G61320, AT2G42850, AT1G11600, AT4G39510, AT5G35715, AT3G19270, AT5G10610, AT2G28850, AT3G25180, AT1G16400, AT2G22330, AT5G04330, AT5G42580, AT5G38970, AT3G53130, AT5G45340, AT3G14690, AT3G48290, AT2G25160, AT1G64900, AT5G42650, AT1G13710, AT2G45570, AT3G53305, AT2G05180, AT3G14610, AT3G52970, AT3G26300, AT1G74540, AT5G04630, AT1G13140, AT4G13290, AT2G26710, AT1G65670, AT5G42590, AT2G45970, AT3G20110, AT1G24540, AT3G30290, AT5G07990, AT5G06900, AT3G26220, AT3G03470, AT4G13310, AT4G13770, AT5G52400, AT1G31800, AT4G37340, AT1G05160, AT4G37370, AT5G63450, AT1G66540, AT3G26310, AT4G12330, AT2G34490, AT1G13150, AT3G20120, AT4G15300, AT2G45580, AT3G20940, AT5G05260, AT2G46960, AT4G12320, AT4G15380, AT4G20240, AT2G16060, AT4G12300, AT4G22690, AT4G36380, AT3G44970, AT5G02900, AT4G00360, AT4G31970, AT3G48320, AT3G14630, AT3G61880, AT3G26125, AT1G19630, AT2G45510, AT4G39500, AT2G23180, AT1G65340, AT3G26320, AT1G13100, AT5G52320, AT3G20130, AT5G36130, AT1G50560, AT4G37330, AT3G20950, AT5G57260, AT5G25900, AT2G42250, AT3G14640, AT4G39480, AT3G53290, AT5G25140, AT4G22710, AT5G58860, AT3G44250, AT3G48280, AT1G63710, AT1G78490, AT4G37430, AT3G26330, AT5G36220, AT3G61040, AT3G30180, AT4G37320, AT5G24950, AT3G26230, AT1G79370, AT2G02580, AT3G48270, AT2G23190
response to auxin (GO:0009733)	AT5G06300, AT1G80680, AT4G37390, AT3G43120, AT5G13370, AT1G74950, AT4G13790, AT2G46690, AT1G19220, AT1G72430, AT2G35940, AT1G19180, AT5G47370, AT1G64520, AT3G06490, AT1G78100, AT2G26740, AT3G12955, AT5G18060, AT3G07390, AT5G27420, AT3G52400, AT5G13300, AT5G59780, AT4G11280, AT5G03310, AT2G24850, AT3G07370, AT4G16420, AT3G28210, AT1G19640, AT2G38120, AT5G37020, AT2G24400, AT3G55120, AT1G59500, AT1G48660, CPR6, AT2G47460, AT1G29510, AT4G34390, AT2G27690, HCA, AT2G21220, AT2G33860, AT5G01270, AT3G09980, AT5G09810, AT4G34710, AT1G59750, AT3G24280, AT4G14550, AT2G25790, AT2G44840, AT2G47260, AT1G29460, AT2G28085, SAR1, AT5G50760, AT5G59430, AT3G49850, AT1G49950, AT1G15520, AT3G28910, AT4G01370, AT5G59220, AT2G39370, AT5G27780, AT5G18020, AT5G20820, AT4G02570, AT4G32810, AT3G25880, AT1G56150, AT1G22920, AT3G14050, AT2G04550, AT5G65940, AT3G15500, AT3G11820, AT1G09540, AT5G20730, AT1G80390, AT5G54500, AT3G01220, AT3G63010, AT5G57560, AT1G74840, AT2G06050, AT4G00080, AT1G56650, AT1G25490, AT5G13930, AT2G45210, AT3G24650, AT1G33410, AT3G09940, AT5G24520, AT2G01570, AT1G52890, AT4G16780, AT5G06960, AT2G46830, AT5G20570, AT1G29420, AT5G05730, AT3G03850, AT3G23050, AT5G02840, AT1G69270, AT5G01490, AT1G75580, AT5G39610, ICR2, AT5G53590, AT4G34760, AT2G31180, AT1G16510, AT2G25930, AT1G04250, AT3G17600, AT4G18010, AT3G61900, AT4G21440, AT3G04730, AT3G15540, AT5G07700, AT1G04100, AT2G02560, AT2G21210, AT1G64060, AT2G46070, AT5G20810, AT1G04550, AT2G46270, AT4G25030, AT3G26760, AT2G34650, AT1G75590, AT3G11260, AT1G66340, AT3G58190, AT1G19850, AT4G38630, AT2G19690, AT2G16580, AT4G32690, AT2G04160, AT5G65670, AT4G34790, AT3G12830, AT4G15430, AT3G20770,

Supplemental Table 4.6 (cont'd)

	AT4G33940, AT1G28480, AT4G34810, AT1G18890, AT1G29500, AT1G52830, AT2G23170, AT1G76190, AT3G09870, AT1G34670, AT5G51470, AT5G12330, AT4G12410, AT3G20830, AT5G37260, AT1G15580, AT1G49010, AT3G47600, AT4G14560, AT3G62980, AT5G13220, AT1G23160, AT4G38850, AT5G57090, AT4G00880, AT1G54060, AT4G01280, AT2G34600, AT1G54100, AT2G22810, AT4G18950, AT3G22650, AT3G16350, AT1G16540, AT5G18030, AT4G03400, AT1G28370, AT2G20000, AT2G35270, AT4G16890, AT1G47510, AT4G31320, AT4G34780, AT3G48360, AT1G61120, AT3G19580, AT1G14000, AT4G26400, AT2G18010, AT2G41820, AT2G34720, AT1G43040, AT5G67580, AT2G36910, AT2G43790, AT5G56290, AT4G05100, AT4G22620, AT4G29940, AT1G29430, AT2G22240, AT1G70000, AT3G03820, AT4G03190, AT4G38860, AT5G42410, AT5G54490, AGE2, AT4G34770, AT1G74100, AT2G46510, AT1G73730, AT5G40770, AT5G55120, AT1G53230, AT1G01560, AT1G04240, AT1G29440, AT5G13350, AT4G17615, AT5G51640, AT1G28130, AT1G73590, AT3G05630, AT2G06850, AT1G17345, AT1G15430, AT3G53250, AT2G37030, AT1G20470, AT5G63160, AT2G39550, AT1G19840, AT3G21110, AT4G18710, AT1G69260, AT2G26710, AT5G67300, AT5G57420, AT4G34800, AT4G23915, AT5G08640, AT1G77690, AT2G21200, AT5G07990, AT1G24590, AT5G25890, AT4G30080, AT1G27740, AT2G25170, AT5G14960, AT5G19140, AT5G64890, AT1G74660, AT3G26090, AT5G15310, GUP1, GUP2, AT4G20380, AT3G59900, AT5G66260, AT4G38840, AT1G48410, AT2G19560, AT2G36210, AT1G09700, AT1G15050, AT5G37770, AT1G17380, AT5G22220, AT3G60690, AT5G45710, AT1G19830, AT3G20220, AT4G36800, AT5G13360, AT4G29080, AT1G29490, AT1G26870, AT1G15750, AT1G08030, AT3G16500, AT2G37630, AT5G18050, AT5G57740, AT1G31340, AT2G47750, AT4G33880, AT5G43700, AT1G63840, AT4G36740, AT2G05710, AT4G27260, AT1G48670, AT1G18570, AT1G22640, AT1G71230, AT3G28860, AT3G62100, AT4G14430, AT4G36110, AT5G17300, AT3G03830, AT1G30330, AT1G10210, AT2G14960, AT2G01200, AT1G32230, AT4G27410, AT2G47190, AT3G57530, AT2G22670, AT1G29450, AT5G10990, AT5G66700, AT1G79130, AT2G46370, AT5G50120, AT5G67480, AT1G01060, AT1G54990, AT4G19690, AT3G02260, AT5G18010, AT4G37295, AT1G17520, AT4G32280, AT4G12550, AT4G32880, AT5G18080, AT4G13520, AT4G09530, AT1G06400, AT5G54510, AT4G28640, AT3G51200, AT4G37610, AT3G23250, AT3G26790, AT1G57560, AT5G13380, AT2G47000, AT3G09600, AT1G51950, AT2G17290, SSA-2, AT3G23030, AT1G48690, AT3G63300, AT3G55730, AT1G27730, AT1G63720, AT4G39403, AT4G26080, AT2G34680, AT2G46990, AT3G03847, AT3G03840, AT2G33310, AT3G50410, AT5G38895, AT4G34750
response to chitin (GO:0010200)	AT5G06300, AT1G80680, AT4G37390, AT3G43120, AT5G13370, AT1G74950, AT4G13790, AT2G46690, AT1G19220, AT1G72430, AT2G35940, AT1G19180, AT5G47370, AT1G64520, AT3G06490, AT1G78100, AT2G26740, AT3G12955, AT5G18060, AT3G07390, AT5G27420, AT3G52400, AT5G13300, AT5G59780, AT4G11280, AT5G03310, AT2G24850, AT3G07370, AT4G16420, AT3G28210, AT1G19640, AT2G38120, AT5G37020, AT2G24400, AT3G55120, AT1G59500, AT1G48660, CPR6, AT2G47460, AT1G29510, AT4G34390, AT2G27690, HCA, AT2G21220, AT2G33860, AT5G01270, AT3G09980, AT5G09810, AT4G34710, AT1G59750, AT3G24280, AT4G14550, AT2G25790, AT2G44840, AT2G47260, AT1G29460, AT2G28085, SAR1, AT5G50760, AT5G59430, AT3G49850, AT1G49950, AT1G15520, AT3G28910, AT4G01370, AT5G59220, AT2G39370, AT5G27780, AT5G18020, AT5G20820, AT4G02570, AT4G32810, AT3G25880, AT1G56150, AT1G22920, AT3G14050, AT2G04550, AT5G65940, AT3G15500, AT3G11820, AT1G09540, AT5G20730, AT1G80390, AT5G54500, AT3G01220, AT3G63010, AT5G57560, AT1G74840, AT2G06050, AT4G00080, AT1G56650, AT1G25490, AT5G13930, AT2G45210, AT3G24650, AT1G33410, AT3G09940, AT5G24520,

Supplemental Table 4.6 (cont'd)

	AT2G01570, AT1G52890, AT4G16780, AT5G06960, AT2G46830, AT5G20570, AT1G29420, AT5G05730, AT3G03850, AT3G23050, AT5G02840, AT1G69270, AT5G01490, AT1G75580, AT5G39610, ICR2, AT5G53590, AT4G34760, AT2G31180, AT1G16510, AT2G25930, AT1G04250, AT3G17600, AT4G18010, AT3G61900, AT4G21440, AT3G04730, AT3G15540, AT5G07700, AT1G04100, AT2G02560, AT2G21210, AT1G64060, AT2G46070, AT5G20810, AT1G04550, AT2G46270, AT4G25030, AT3G26760, AT2G34650, AT1G75590, AT3G11260, AT1G66340, AT3G58190, AT1G19850, AT4G38630, AT2G19690, AT2G16580, AT4G32690, AT2G04160, AT5G65670, AT4G34790, AT3G12830, AT4G15430, AT3G20770, AT4G33940, AT1G28480, AT4G34810, AT1G18890, AT1G29500, AT1G52830, AT2G23170, AT1G76190, AT3G09870, AT1G34670, AT5G51470, AT5G12330, AT4G12410, AT3G20830, AT5G37260, AT1G15580, AT1G49010, AT3G47600, AT4G14560, AT3G62980, AT5G13220, AT1G23160, AT4G38850, AT5G57090, AT4G00880, AT1G54060, AT4G01280, AT2G34600, AT1G54100, AT2G22810, AT4G18950, AT3G22650, AT3G16350, AT1G16540, AT5G18030, AT4G03400, AT1G28370, AT2G20000, AT2G35270, AT4G16890, AT1G47510, AT4G31320, AT4G34780, AT3G48360, AT1G61120, AT3G19580, AT1G14000, AT4G26400, AT2G18010, AT2G41820, AT2G34720, AT1G43040, AT5G67580, AT2G36910, AT2G43790, AT5G56290, AT4G05100, AT4G22620, AT4G29940, AT1G29430, AT2G22240, AT1G70000, AT3G03820, AT4G03190, AT4G38860, AT5G42410, AT5G54490, AGE2, AT4G34770, AT1G74100, AT2G46510, AT1G73730, AT5G40770, AT5G55120, AT1G53230, AT1G01560, AT1G04240, AT1G29440, AT5G13350, AT4G17615, AT5G51640, AT1G28130, AT1G73590, AT3G05630, AT2G06850, AT1G17345, AT1G15430, AT3G53250, AT2G37030, AT1G20470, AT5G63160, AT2G39550, AT1G19840, AT3G21110, AT4G18710, AT1G69260, AT2G26710, AT5G67300, AT5G57420, AT4G34800, AT4G23915, AT5G08640, AT1G77690, AT2G21200, AT5G07990, AT1G24590, AT5G25890, AT4G30080, AT1G27740, AT2G25170, AT5G14960, AT5G19140, AT5G64890, AT1G74660, AT3G26090, AT5G15310, GUP1, GUP2, AT4G20380, AT3G59900, AT5G66260, AT4G38840, AT1G48410, AT2G19560, AT2G36210, AT1G09700, AT1G15050, AT5G37770, AT1G17380, AT5G22220, AT3G60690, AT5G45710, AT1G19830, AT3G20220, AT4G36800, AT5G13360, AT4G29080, AT1G29490, AT1G26870, AT1G15750, AT1G08030, AT3G16500, AT2G37630, AT5G18050, AT5G57740, AT1G31340, AT2G47750, AT4G33880, AT5G43700, AT1G63840, AT4G36740, AT2G05710, AT4G27260, AT1G48670, AT1G18570, AT1G22640, AT1G71230, AT3G28860, AT3G62100, AT4G14430, AT4G36110, AT5G17300, AT3G03830, AT1G30330, AT1G10210, AT2G14960, AT2G01200, AT1G32230, AT4G27410, AT2G47190, AT3G57530, AT2G22670, AT1G29450, AT5G10990, AT5G66700, AT1G79130, AT2G46370, AT5G50120, AT5G67480, AT1G01060, AT1G54990, AT4G19690, AT3G02260, AT5G18010, AT4G37295, AT1G17520, AT4G32280, AT4G12550, AT4G32880, AT5G18080, AT4G13520, AT4G09530, AT1G06400, AT5G54510, AT4G28640, AT3G51200, AT4G37610, AT3G23250, AT3G26790, AT1G57560, AT5G13380, AT2G47000, AT3G09600, AT1G51950, AT2G17290, SSA-2, AT3G23030, AT1G48690, AT3G63300, AT3G55730, AT1G27730, AT1G63720, AT4G39403, AT4G26080, AT2G34680, AT2G46990, AT3G03847, AT3G03840, AT2G33310, AT3G50410, AT5G38895, AT4G34750
RNA binding (GO:0003723)	AT3G50670, AT4G26650, AT1G49760, AT5G43110, AT1G14640, AT3G10360, AT5G18110, AT5G55670, AT4G36690, AT4G16830, AT3G55460, AT2G31890, AT1G67770, AT2G42890, AT4G13860, AT2G43640, AT3G20250, AT2G47220, AT2G14870, AT2G07734, AT5G19350, AT5G18810, AT4G29090, AT1G71720, AT4G22380, AT3G20890, AT4G34110, AT1G28090, AT1G27650, AT5G55100, AT1G18050, AT3G55340, AT5G50250, AT1G07350, AT1G35730, AT5G19030, AT3G11400, AT1G60000, AT2G47580, AT5G47210, AT1G64810, AT4G09040,

Supplemental Table 4.6 (cont'd)

AT1G74350, AT3G07250, AT2G19380, AT3G26120, AT1G74230, AT5G60170, AT4G27000, AT5G12190, AT3G19130, AT4G27490, AT4G02430, AT3G54770, AT3G10845, AT5G07060, AT3G07810, AT3G51950, AT1G60650, AT5G54900, AT3G61620, AT3G63450, AT1G53120, ATCG00040, AT3G27700, AT3G02320, AT1G33470, AT3G20550, AT1G58470, AT5G60960, AT3G56860, AT1G09340, AT1G55310, AT1G80930, AT1G22760, AT1G50300, AT3G27750, AT2G30260, AT4G24270, AT4G10110, AT4G26370, AT3G49430, AT1G49600, AT1G17640, AT3G45630, AT3G49390, AT5G04210, AT3G15010, AT5G59860, AT3G26560, AT5G61780, AT5G55550, AT3G46020, AT3G12680, AT4G17610, AT3G23830, AT1G06960, AT5G23690, AT3G52150, AT1G34140, AT2G37510, AT2G22090, AT1G13190, AT1G47490, AT3G56330, AT4G25500, AT1G62670, AT5G14580, AT1G09140, AT1G01760, AT2G05160, AT4G08840, AT3G55280, AT1G09230, AT1G22240, AT4G25630, AT2G29190, AT5G06520, AT1G13730, AT5G18180, AT5G51120, AT3G16810, AT1G12800, AT1G14650, AT2G33410, AT2G29580, AT5G52040, AT1G33520, AT4G03110, AT5G60110, AT2G19870, AT5G58130, AT5G07350, AT5G08620, AT5G10350, AT5G16260, AT5G42820, AT4G12600, AT5G09610, AT5G11530, AT3G48830, AT5G24440, AT1G29590, AT2G47310, AT1G56030, AT3G06970, AT5G19960, AT1G07360, AT4G25880, AT3G03710, AT2G29200, AT1G24450, AT3G19090, AT3G13570, AT2G23350, AT1G15910, AT4G11175, AT1G43860, AT2G25900, AT1G16610, AT4G10610, AT5G53180, AT5G53720, AT1G30460, AT3G13224, AT5G02250, AT4G15520, AT1G76460, AT1G29550, AT1G30010, AT3G14450, AT5G44200, AT3G53460, AT3G21215, AT2G18510, AT1G53720, AT5G46250, AT1G09150, AT1G22910, AT5G25060, AT4G38020, AT1G53650, AT4G19610, AT2G21690, AT2G03640, AT1G60830, AT5G48870, AT3G16380, AT1G37140, AT5G64200, AT2G33435, AT3G03920, AT3G43920, AT3G08010, AT2G39460, AT2G21660, AT5G05720, AT2G43960, AT5G04600, AT1G01080, AT4G16280, AT5G57870, AT4G36020, AT5G47620, AT3G52120, AT3G29390, AT1G71800, AT5G30510, AT3G21100, AT2G29140, AT4G17520, AT3G22310, AT2G43410, AT3G23700, AT4G15417, AT3G26420, AT3G10400, AT1G45100, AT2G24350, AT4G16200, AT1G21620, AT1G18630, AT1G43190, AT1G71770, AT3G61860, AT4G24770, AT4G34730, AT4G39260, AT2G46780, AT1G78260, AT2G39120, AT4G14300, AT4G12640, AT2G33440, AT1G60080, AT3G25150, AT5G61030, AT4G20030, AT2G39260, AT5G41690, AT3G08000, AT1G77680, AT2G21440, AT5G66010, AT5G07290, AT2G39140, AT1G22330, AT3G01150, AT5G26880, AT5G60180, AT1G29400, AT5G04280, AT2G41060, AT5G52490, AT3G46210, AT5G64390, AT5G20320, AT2G22100, AT4G18120, AT5G15390, AT5G65260, AT4G18040, AT5G43090, AT1G35750, AT5G28390, AT4G00830, AT2G17510, AT1G69250, AT2G36660, AT5G06210, AT1G73490, AT2G37220, AT2G43370, AT3G47120, AT3G07750, AT5G23080, AT1G02840, AT5G10800, AT5G48650, AT5G08180, AT2G44710, AT1G51510, AT5G15750, AT4G31200, AT2G46610, AT5G54580, AT3G12990, AT5G60980, AT3G13700, AT1G78160, AT1G73530, AT3G04500, AT5G58040, AT5G43960, AT3G60500, AT5G59280, AT5G15810, AT4G13850, AT4G36960, AT2G35410, AT5G56510, AT5G06000, AT1G11650, AT1G72320, AT1G20880, AT1G35850, AT4G32720, AT2G24050, AT5G51410, AT2G20490, AT3G12640, AT5G61960, AT3G11964, AT1G10320, AT3G52660, AT3G20930, AT1G47500, AT5G20160, AT4G37510, AT5G35620, AT2G17580, AT1G32790, AT2G40290, AT5G40490, AT5G47320, AT3G52380, AT1G60900, AT5G46920, AT5G03580, AT5G53680, AT5G05470, AT5G46840, AT1G03457

Supplemental Table 4.6 (cont'd)

transferase activity, transferring glycosyl groups (GO:0016757)	AT2G37730, AT4G02500, AT5G15740, AT2G29740, AT1G11720, AT5G05870, AT2G31750, AT2G04560, AT3G46720, AT2G03280, AT3G46680, AT4G26250, AT4G00300, AT1G60450, AT1G74420, AT4G03340, AT1G09350, AT5G20410, AT1G54940, AT1G67850, AT5G64740, AT3G21780, AT4G15290, AT1G52420, AT2G15350, AT4G19900, AT4G13410, AT1G22370, AT1G32930, AT3G27470, AT3G59100, AT1G71220, AT5G13500, AT3G61130, AT5G41460, AT1G30530, AT4G04970, AT2G32540, AT3G46970, AT5G03490, AT2G25540, AT1G13250, AT1G75420, AT2G44660, AT3G46650, AT2G31960, AT5G14860, AT1G27600, AT2G22930, AT1G70090, AT4G18780, AT3G26370, AT1G38131, AT1G24570, AT4G15320, AT2G30140, AT2G28080, AT4G17770, AT3G02250, AT1G24170, AT5G37200, AT1G24070, AT3G11540, AT4G16600, AT2G03220, AT4G15550, AT2G24630, AT1G20550, AT4G03550, AT3G62660, AT2G32430, AT1G08660, AT3G29320, AT1G68380, AT1G73810, AT5G01100, AT1G14070, AT5G16190, AT4G21060, AT2G26100, AT2G19160, AT5G38010, AT3G16520, AT5G59530, AT1G68020, AT5G49690, AT1G77810, AT1G50580, AT2G21770, AT3G07020, AT2G23260, AT4G36890, AT5G04480, AT1G78280, AT3G53160, AT1G70290, AT5G53340, AT2G35650, AT5G12260, AT2G35710, AT2G37980, AT4G09500, AT2G32610, AT4G14100, AT1G04920, AT5G17040, AT5G47780, AT3G02350, AT3G21310, AT3G01040, AT1G01570, AT1G11730, AT5G15050, AT3G50760, AT5G01220, AT4G19460, AT1G70630, AT5G17030, AT1G28240, AT1G38065, AT1G24100, AT4G12700, AT1G34550, AT1G05570, AT3G21770, AT4G15500, AT1G74800, AT1G29200, AT3G57200, AT3G01720, AT1G80290, AT1G16900, AT1G23480, AT1G05150, AT4G11350, AT5G65470, AT1G22340, AT5G11730, AT1G60140, AT2G20810, AT1G77130, AT4G32110, AT2G36750, AT2G26480, AT5G03760, AT5G09870, AT3G52060, AT3G03810, AT1G10880, AT1G08040, AT3G21190, AT1G22460, AT4G39350, AT2G16890, AT4G16590, AT2G43820, AT3G54100, AT1G67880, AT5G49190, AT1G35510, AT1G11940, AT5G15470, AT3G43190, AT4G24000, AT2G30150, AT1G26810, AT2G02910, AT1G74380, AT3G42180, AT2G31790, AT3G57420, AT1G04910, AT1G62330, AT2G01480, AT2G36780, AT2G03210, AT4G15260, AT5G57270, AT4G26940, AT5G05880, AT1G71070, AT1G68390, AT1G08990, AT3G11420, AT4G02280, AT2G44500, AT3G11670, AT3G01620, AT5G46220, AT3G03690, AT4G33330, AT4G32410, AT5G13000, AT1G18690, AT3G02100, AT3G06440, AT4G31350, AT1G61050, AT1G05530, AT2G15390, AT2G23210, AT2G30575, AT1G71990, AT4G38310, AT3G56000, AT5G54010, AT5G25330, AT4G12840, AT5G57500, AT3G25140, AT5G38460, AT5G03770, AT4G36770, AT5G05170, AT2G41451, AT5G59070, AT4G27550, AT1G73740, AT5G60700, AT1G78580, AT2G41150, AT1G52630, AT3G28180, AT1G14020, AT2G29730, AT5G01250, AT4G02130, AT1G11990, AT3G46700, AT5G22740, AT1G16980, AT1G22400, AT3G14570, AT1G60470, AT1G63450, AT4G25870, AT3G14960, AT1G05560, AT1G02720, AT2G37580, AT2G15480, AT1G51630, AT3G45100, AT5G64600, AT3G21760, AT5G63390, AT1G73160, AT1G20570, AT1G53290, AT5G42660, AT1G07240, AT4G32290, AT3G29630, AT5G59520, AT3G07170, AT3G30300, AT4G38190, AT1G05280, AT1G43620, AT1G23870, AT3G08550, AT1G06410, AT4G37690, AT1G03520, AT1G06780, AT1G32180, AT1G01420, AT5G18480, AT2G36760, AT4G38270, AT3G55700, AT5G25970, AT3G46670, AT3G24040, AT5G22130, AT4G09630, AT2G43840, AT2G15370, AT3G15350, AT2G22590, AT2G28310, AT5G54060, AT1G64910, AT2G18560, AT1G11170, AT2G38650, AT2G38150, AT3G15940, AT4G27560, AT5G23790, AT5G35570, AT5G44030, AT4G24530, AT4G24010, AT2G29750, AT2G25300, AT1G73880, AT5G24300, AT3G58790, AT1G06000, AT4G10120, AT1G76270, AT4G01070, AT1G62305, AT1G55850, AT1G78800, AT3G19280, AT4G15270, AT3G62720, AT2G36790, AT1G16570, AT2G33100, AT5G05860, AT1G01390, AT2G32450, AT2G13290, AT3G46690, AT2G36850,

Supplemental Table 4.6 (cont'd)

	AT3G21800, AT5G05890, AT2G36970, AT4G16710, AT1G14970, AT1G07260, AT4G16650, AT1G22380, AT4G15280, AT3G03050, AT1G14080, AT4G00550, AT4G23490, AT4G31590, AT4G23990, AT1G51770, AT4G38240, AT5G22070, AT1G10280, AT4G17430, AT1G22360, AT2G35100, AT1G33430, AT1G07850, AT3G06260, AT5G25990, AT1G53040, AT5G67230, AT4G38390, AT1G56600, AT2G37090, AT3G55830, AT4G27480, AT4G18240, AT1G61240, AT4G18530, AT2G22900, AT2G39630, AT5G14850, AT1G64920, AT2G20370, AT5G04500, AT3G50740, AT3G09020, AT3G01180, AT2G46480, AT5G05900, AT1G33250, AT2G04280, AT5G53990, AT5G07720, AT5G25265, AT4G30060, AT3G56750, AT4G15480, AT2G15490, AT2G23250, AT3G48820, AT4G31780, AT5G14550, AT3G55710, AT4G15240, AT1G05680, AT3G21750, AT4G07960, AT1G07250, AT5G14480, AT1G19300, AT2G36800, AT1G05170, AT1G19710, AT1G05670, AT4G01210, AT1G32900, AT5G66690, AT2G41770, AT5G65550, AT1G10400, AT3G27540, AT5G54690, AT3G22250, AT2G36770, AT1G06490, AT4G38500, AT5G26310, AT2G32530, AT5G16170, AT3G10630, AT3G46660, AT1G12990, AT2G40190, AT3G07900, AT4G15490, AT2G25260, AT1G51210, AT3G07330, AT2G18570, AT3G18660, AT1G49710, AT3G04240, AT1G13000, AT2G32620, AT4G27570, AT5G17050, AT5G16910, AT1G27120, AT5G20280, AT4G08810, AT2G18700, AT3G26440, AT2G29710, AT5G12890
translation (GO:0006412)	AT1G14400, AT3G07550, AT3G21860, AT2G03170, AT3G09770, AT2G16920, AT2G33770, AT2G36370, AT3G06140, AT5G02750, AT2G26000, AT1G80570, AT4G03510, AT2G01150, AT3G54780, AT4G27470, AT3G29270, AT1G10230, AT5G06460, AT4G33160, AT1G63900, AT5G43190, AT3G08690, AT2G02760, AT1G12820, AT2G32950, AT4G34210, AT4G04690, AT1G02860, AT5G49980, AT1G51550, AT5G59300, AT4G12570, AT5G46210, AT4G10160, AT3G60220, AT1G36340, AT3G53060, AT2G30110, AT3G26810, AT1G49210, AT4G19700, AT2G44950, AT5G14420, AT3G13550, AT5G27420, AT3G52560, AT2G04660, AT4G37890, AT2G30580, AT1G75950, AT4G34470, AT5G05280, AT5G10380, AT5G65683, AT1G50490, AT1G55860, AT3G60020, AT1G78870, AT5G42190, AT5G05560, AT2G44330, AT4G03190, AT3G61590, AT4G25230, AT2G22010, AT5G60710, AT3G21840, AT3G46620, AT3G07370, AT3G19140, AT2G42620, AT4G28370, AT2G16740, AT5G50430, AT5G02310, AT4G11360, AT1G79810, AT3G55530, AT3G05870, AT1G77000, AT4G05470, AT5G42200, AT1G57820, AT5G27920, AT5G51450, AT3G60010, AT5G41700, AT3G17205, AT3G12630, AT5G38070, AT3G25650, AT2G28830, AT1G75440, AT4G14220, AT5G53300, AT5G50870, AT3G05545, AT5G02920, AT3G07360, AT1G20140, AT4G27960, AT2G16810, AT4G05460, AT3G08700, AT3G24515, AT2G39810, AT2G45950, AT2G35000, AT1G64230, AT3G54850, AT2G38970, AT2G04920, AT2G32790, AT1G66050, AT1G08050, AT2G25700, AT1G53020, AT3G53410, AT3G04460, AT1G26830, AT5G67250, AT4G05490, AT1G65430, AT1G30950, AT3G42830, AT1G69670, AT4G07400, AT3G53090, AT1G70320, AT1G23260, AT3G61415, AT3G55380, AT4G08980, AT3G17000, AT1G70660, AT5G56150, AT1G29340, AT5G05080, AT5G25760, AT5G45100, AT1G79380, AT1G45050, AT4G36410, AT5G57740, AT1G20780, AT2G39940, AT4G23450, AT2G25490, AT1G27910, AT4G30640, AT5G53840, AT1G22500, AT1G17280, AT3G24800, AT4G28270, AT1G06770, AT5G49665, AT3G21850, AT5G02880, AT5G18650, AT5G03200, AT1G10560, AT1G63800, AT5G57360, AT5G42990, SFO1, AT2G18600, AT2G46030, AT4G33210, AT5G13530, AT3G20060, AT2G35930, AT2G26350, AT3G47990, AT5G39550, AT2G21950, AT1G15100, AT2G22680, AT1G76920, AT3G18710, AT4G28890, AT3G01650, AT5G07270, AT3G50080, AT2G18915, AT5G20910, AT2G42360, AT5G19080, AT5G41340, AT2G36060, AT2G44900, AT4G21070, AT3G60350, AT2G20160, AT3G46460, AT1G14260, AT3G12775, AT3G05200, AT3G56580, AT2G42160, AT1G51290, AT1G16890, AT3G21830, AT4G24390,

Supplemental Table 4.6 (cont'd)

	AT5G62540, AT2G47700, AT2G03160, AT3G54650, AT5G65200, AT3G21410, AT3G63530, AT3G06330, AT5G59550, AT5G63970, AT5G53910, AT3G23280, AT1G21410, AT2G03190, AT3G52450, AT1G12760, AT1G51320, AT1G47056, AT1G68050, AT3G46510, AT4G15475, AT2G20650, AT3G62980, AT2G03560, AT5G01720, AT4G08590, AT4G38600, AT2G17020, AT4G02440, AT4G34100
transporter activity (GO:0005215)	AT2G37730, AT4G02500, AT5G15740, AT2G29740, AT1G11720, AT5G05870, AT2G31750, AT2G04560, AT3G46720, AT2G03280, AT3G46680, AT4G26250, AT4G00300, AT1G60450, AT1G74420, AT4G03340, AT1G09350, AT5G20410, AT1G54940, AT1G67850, AT5G64740, AT3G21780, AT4G15290, AT1G52420, AT2G15350, AT4G19900, AT4G13410, AT1G22370, AT1G32930, AT3G27470, AT3G59100, AT1G71220, AT5G13500, AT3G61130, AT5G41460, AT1G30530, AT4G04970, AT2G32540, AT3G46970, AT5G03490, AT2G25540, AT1G13250, AT1G75420, AT2G44660, AT3G46650, AT2G31960, AT5G14860, AT1G27600, AT2G22930, AT1G70090, AT4G18780, AT3G26370, AT1G38131, AT1G24570, AT4G15320, AT2G30140, AT2G28080, AT4G17770, AT3G02250, AT1G24170, AT5G37200, AT1G24070, AT3G11540, AT4G16600, AT2G03220, AT4G15550, AT2G24630, AT1G20550, AT4G03550, AT3G62660, AT2G32430, AT1G08660, AT3G29320, AT1G68380, AT1G73810, AT5G01100, AT1G14070, AT5G16190, AT4G21060, AT2G26100, AT2G19160, AT5G38010, AT3G16520, AT5G59530, AT1G68020, AT5G49690, AT1G77810, AT1G50580, AT2G21770, AT3G07020, AT2G23260, AT4G36890, AT5G04480, AT1G78280, AT3G53160, AT1G70290, AT5G53340, AT2G35650, AT5G12260, AT2G35710, AT2G37980, AT4G09500, AT2G32610, AT4G14100, AT1G04920, AT5G17040, AT5G47780, AT3G02350, AT3G21310, AT3G01040, AT1G01570, AT1G11730, AT5G15050, AT3G50760, AT5G01220, AT4G19460, AT1G70630, AT5G17030, AT1G28240, AT1G38065, AT1G24100, AT4G12700, AT1G34550, AT1G05570, AT3G21770, AT4G15500, AT1G74800, AT1G29200, AT3G57200, AT3G01720, AT1G80290, AT1G16900, AT1G23480, AT1G05150, AT4G11350, AT5G65470, AT1G22340, AT5G11730, AT1G60140, AT2G20810, AT1G77130, AT4G32110, AT2G36750, AT2G26480, AT5G03760, AT5G09870, AT3G52060, AT3G03810, AT1G10880, AT1G08040, AT3G21190, AT1G22460, AT4G39350, AT2G16890, AT4G16590, AT2G43820, AT3G54100, AT1G67880, AT5G49190, AT1G35510, AT1G11940, AT5G15470, AT3G43190, AT4G24000, AT2G30150, AT1G26810, AT2G02910, AT1G74380, AT3G42180, AT2G31790, AT3G57420, AT1G04910, AT1G62330, AT2G01480, AT2G36780, AT2G03210, AT4G15260, AT5G57270, AT4G26940, AT5G05880, AT1G71070, AT1G68390, AT1G08990, AT3G11420, AT4G02280, AT2G44500, AT3G11670, AT3G01620, AT5G46220, AT3G03690, AT4G33330, AT4G32410, AT5G13000, AT1G18690, AT3G02100, AT3G06440, AT4G31350, AT1G61050, AT1G05530, AT2G15390, AT2G23210, AT2G30575, AT1G71990, AT4G38310, AT3G56000, AT5G54010, AT5G25330, AT4G12840, AT5G57500, AT3G25140, AT5G38460, AT5G03770, AT4G36770, AT5G05170, AT2G41451, AT5G59070, AT4G27550, AT1G73740, AT5G60700, AT1G78580, AT2G41150, AT1G52630, AT3G28180, AT1G14020, AT2G29730, AT5G01250, AT4G02130, AT1G11990, AT3G46700, AT5G22740, AT1G16980, AT1G22400, AT3G14570, AT1G60470, AT1G63450, AT4G25870, AT3G14960, AT1G05560, AT1G02720, AT2G37580, AT2G15480, AT1G51630, AT3G45100, AT5G64600, AT3G21760, AT5G63390, AT1G73160, AT1G20570, AT1G53290, AT5G42660, AT1G07240, AT4G32290, AT3G29630, AT5G59520, AT3G07170, AT3G30300, AT4G38190, AT1G05280, AT1G43620, AT1G23870, AT3G08550, AT1G06410, AT4G37690, AT1G03520, AT1G06780, AT1G32180, AT1G01420, AT5G18480, AT2G36760, AT4G38270, AT3G55700, AT5G25970, AT3G46670, AT3G24040, AT5G22130, AT4G09630, AT2G43840, AT2G15370, AT3G15350, AT2G22590, AT2G28310, AT5G54060, AT1G64910, AT2G18560, AT1G11170, AT2G38650, AT2G38150, AT3G15940,

Supplemental Table 4.6 (cont'd)

	AT4G27560, AT5G23790, AT5G35570, AT5G44030, AT4G24530, AT4G24010, AT2G29750, AT2G25300, AT1G73880, AT5G24300, AT3G58790, AT1G06000, AT4G10120, AT1G76270, AT4G01070, AT1G62305, AT1G55850, AT1G78800, AT3G19280, AT4G15270, AT3G62720, AT2G36790, AT1G16570, AT2G33100, AT5G05860, AT1G01390, AT2G32450, AT2G13290, AT3G46690, AT2G36850, AT3G21800, AT5G05890, AT2G36970, AT4G16710, AT1G14970, AT1G07260, AT4G16650, AT1G22380, AT4G15280, AT3G03050, AT1G14080, AT4G00550, AT4G23490, AT4G31590, AT4G23990, AT1G51770, AT4G38240, AT5G22070, AT1G10280, AT4G17430, AT1G22360, AT2G35100, AT1G33430, AT1G07850, AT3G06260, AT5G25990, AT1G53040, AT5G67230, AT4G38390, AT1G56600, AT2G37090, AT3G55830, AT4G27480, AT4G18240, AT1G61240, AT4G18530, AT2G22900, AT2G39630, AT5G14850, AT1G64920, AT2G20370, AT5G04500, AT3G50740, AT3G09020, AT3G01180, AT2G46480, AT5G05900, AT1G33250, AT2G04280, AT5G53990, AT5G07720, AT5G25265, AT4G30060, AT3G56750, AT4G15480, AT2G15490, AT2G23250, AT3G48820, AT4G31780, AT5G14550, AT3G55710, AT4G15240, AT1G05680, AT3G21750, AT4G07960, AT1G07250, AT5G14480, AT1G19300, AT2G36800, AT1G05170, AT1G19710, AT1G05670, AT4G01210, AT1G32900, AT5G66690, AT2G41770, AT5G65550, AT1G10400, AT3G27540, AT5G54690, AT3G22250, AT2G36770, AT1G06490, AT4G38500, AT5G26310, AT2G32530, AT5G16170, AT3G10630, AT3G46660, AT1G12990, AT2G40190, AT3G07900, AT4G15490, AT2G25260, AT1G51210, AT3G07330, AT2G18570, AT3G18660, AT1G49710, AT3G04240, AT1G13000, AT2G32620, AT4G27570, AT5G17050, AT5G16910, AT1G27120, AT5G20280, AT4G08810, AT2G18700, AT3G26440, AT2G29710, AT5G12890
ubiquitin-protein transferase activity (GO:0004842)	AT1G14400, AT3G07550, AT3G21860, AT2G03170, AT3G09770, AT2G16920, AT2G33770, AT2G36370, AT3G06140, AT5G02750, AT2G26000, AT1G80570, AT4G03510, AT2G01150, AT3G54780, AT4G27470, AT3G29270, AT1G10230, AT5G06460, AT4G33160, AT1G63900, AT5G43190, AT3G08690, AT2G02760, AT1G12820, AT2G32950, AT4G34210, AT4G04690, AT1G02860, AT5G49980, AT1G51550, AT5G59300, AT4G12570, AT5G46210, AT4G10160, AT3G60220, AT1G36340, AT3G53060, AT2G30110, AT3G26810, AT1G49210, AT4G19700, AT2G44950, AT5G14420, AT3G13550, AT5G27420, AT3G52560, AT2G04660, AT4G37890, AT2G30580, AT1G75950, AT4G34470, AT5G05280, AT5G10380, AT5G65683, AT1G50490, AT1G55860, AT3G60020, AT1G78870, AT5G42190, AT5G05560, AT2G44330, AT4G03190, AT3G61590, AT4G25230, AT2G22010, AT5G60710, AT3G21840, AT3G46620, AT3G07370, AT3G19140, AT2G42620, AT4G28370, AT2G16740, AT5G50430, AT5G02310, AT4G11360, AT1G79810, AT3G55530, AT3G05870, AT1G77000, AT4G05470, AT5G42200, AT1G57820, AT5G27920, AT5G51450, AT3G60010, AT5G41700, AT3G17205, AT3G12630, AT5G38070, AT3G25650, AT2G28830, AT1G75440, AT4G14220, AT5G53300, AT5G50870, AT3G05545, AT5G02920, AT3G07360, AT1G20140, AT4G27960, AT2G16810, AT4G05460, AT3G08700, AT3G24515, AT2G39810, AT2G45950, AT2G35000, AT1G64230, AT3G54850, AT2G38970, AT2G04920, AT2G32790, AT1G66050, AT1G08050, AT2G25700, AT1G53020, AT3G53410, AT3G04460, AT1G26830, AT5G67250, AT4G05490, AT1G65430, AT1G30950, AT3G42830, AT1G69670, AT4G07400, AT3G53090, AT1G70320, AT1G23260, AT3G61415, AT3G55380, AT4G08980, AT3G17000, AT1G70660, AT5G56150, AT1G29340, AT5G05080, AT5G25760, AT5G45100, AT1G79380, AT1G45050, AT4G36410, AT5G57740, AT1G20780, AT2G39940, AT4G23450, AT2G25490, AT1G27910, AT4G30640, AT5G53840, AT1G22500, AT1G17280, AT3G24800, AT4G28270, AT1G06770, AT5G49665, AT3G21850, AT5G02880, AT5G18650, AT5G03200, AT1G10560, AT1G63800, AT5G57360, AT5G42990, SFO1, AT2G18600, AT2G46030, AT4G33210, AT5G13530, AT3G20060, AT2G35930, AT2G26350, AT3G47990,

Supplemental Table 4.6 (cont'd)

	AT5G39550, AT2G21950, AT1G15100, AT2G22680, AT1G76920, AT3G18710, AT4G28890, AT3G01650, AT5G07270, AT3G50080, AT2G18915, AT5G20910, AT2G42360, AT5G19080, AT5G41340, AT2G36060, AT2G44900, AT4G21070, AT3G60350, AT2G20160, AT3G46460, AT1G14260, AT3G12775, AT3G05200, AT3G56580, AT2G42160, AT1G51290, AT1G16890, AT3G21830, AT4G24390, AT5G62540, AT2G47700, AT2G03160, AT3G54650, AT5G65200, AT3G21410, AT3G63530, AT3G06330, AT5G59550, AT5G63970, AT5G53910, AT3G23280, AT1G21410, AT2G03190, AT3G52450, AT1G12760, AT1G51320, AT1G47056, AT1G68050, AT3G46510, AT4G15475, AT2G20650, AT3G62980, AT2G03560, AT5G01720, AT4G08590, AT4G38600, AT2G17020, AT4G02440, AT4G34100
zinc ion binding (GO:0008270)	AT1G14400, AT3G07550, AT3G21860, AT2G03170, AT3G09770, AT2G16920, AT2G33770, AT2G36370, AT3G06140, AT5G02750, AT2G26000, AT1G80570, AT4G03510, AT2G01150, AT3G54780, AT4G27470, AT3G29270, AT1G10230, AT5G06460, AT4G33160, AT1G63900, AT5G43190, AT3G08690, AT2G02760, AT1G12820, AT2G32950, AT4G34210, AT4G04690, AT1G02860, AT5G49980, AT1G51550, AT5G59300, AT4G12570, AT5G46210, AT4G10160, AT3G60220, AT1G36340, AT3G53060, AT2G30110, AT3G26810, AT1G49210, AT4G19700, AT2G44950, AT5G14420, AT3G13550, AT5G27420, AT3G52560, AT2G04660, AT4G37890, AT2G30580, AT1G75950, AT4G34470, AT5G05280, AT5G10380, AT5G65683, AT1G50490, AT1G55860, AT3G60020, AT1G78870, AT5G42190, AT5G05560, AT2G44330, AT4G03190, AT3G61590, AT4G25230, AT2G22010, AT5G60710, AT3G21840, AT3G46620, AT3G07370, AT3G19140, AT2G42620, AT4G28370, AT2G16740, AT5G50430, AT5G02310, AT4G11360, AT1G79810, AT3G55530, AT3G05870, AT1G77000, AT4G05470, AT5G42200, AT1G57820, AT5G27920, AT5G51450, AT3G60010, AT5G41700, AT3G17205, AT3G12630, AT5G38070, AT3G25650, AT2G28830, AT1G75440, AT4G14220, AT5G53300, AT5G50870, AT3G05545, AT5G02920, AT3G07360, AT1G20140, AT4G27960, AT2G16810, AT4G05460, AT3G08700, AT3G24515, AT2G39810, AT2G45950, AT2G35000, AT1G64230, AT3G54850, AT2G38970, AT2G04920, AT2G32790, AT1G66050, AT1G08050, AT2G25700, AT1G53020, AT3G53410, AT3G04460, AT1G26830, AT5G67250, AT4G05490, AT1G65430, AT1G30950, AT3G42830, AT1G69670, AT4G07400, AT3G53090, AT1G70320, AT1G23260, AT3G61415, AT3G55380, AT4G08980, AT3G17000, AT1G70660, AT5G56150, AT1G29340, AT5G05080, AT5G25760, AT5G45100, AT1G79380, AT1G45050, AT4G36410, AT5G57740, AT1G20780, AT2G39940, AT4G23450, AT2G25490, AT1G27910, AT4G30640, AT5G53840, AT1G22500, AT1G17280, AT3G24800, AT4G28270, AT1G06770, AT5G49665, AT3G21850, AT5G02880, AT5G18650, AT5G03200, AT1G10560, AT1G63800, AT5G57360, AT5G42990, SFO1, AT2G18600, AT2G46030, AT4G33210, AT5G13530, AT3G20060, AT2G35930, AT2G26350, AT3G47990, AT5G39550, AT2G21950, AT1G15100, AT2G22680, AT1G76920, AT3G18710, AT4G28890, AT3G01650, AT5G07270, AT3G50080, AT2G18915, AT5G20910, AT2G42360, AT5G19080, AT5G41340, AT2G36060, AT2G44900, AT4G21070, AT3G60350, AT2G20160, AT3G46460, AT1G14260, AT3G12775, AT3G05200, AT3G56580, AT2G42160, AT1G51290, AT1G16890, AT3G21830, AT4G24390, AT5G62540, AT2G47700, AT2G03160, AT3G54650, AT5G65200, AT3G21410, AT3G63530, AT3G06330, AT5G59550, AT5G63970, AT5G53910, AT3G23280, AT1G21410, AT2G03190, AT3G52450, AT1G12760, AT1G51320, AT1G47056, AT1G68050, AT3G46510, AT4G15475, AT2G20650, AT3G62980, AT2G03560, AT5G01720, AT4G08590, AT4G38600, AT2G17020, AT4G02440, AT4G34100

Supplemental Table 4.7 TF genes belonging to each TF family in *A. thaliana*

TF Family	TFs
AP2	AT1G64380, AT5G47220, AT3G20310, AT4G18450, AT2G35700, AT4G28140, AT5G57390, AT1G25560, AT1G22810, AT2G20350, AT1G12610, AT1G71520, AT4G25480, AT2G23340, AT1G46768, AT3G25730, AT5G65510, AT2G40340, AT5G65130, AT5G18450, AT5G21960, AT5G61600, AT1G36060, AT4G25470, AT1G21910, AT4G36900, AT1G79700, AT4G39780, AT3G50260, AT1G80580, AT1G28160, AT4G11140, AT1G12890, AT5G53290, AT3G61630, AT5G64750, AT5G25190, AT3G60490, AT5G51990, AT1G12630, AT2G33710, AT2G44940, AT5G67000, AT2G39250, AT1G71130, AT1G06160, AT3G16770, AT1G51120, AT5G17430, AT1G72570, AT1G04370, AT2G31230, AT1G19210, AT3G54990, AT2G36450, AT5G61890, AT5G44210, AT1G50640, AT4G34410, AT5G43410, AT1G77640, AT4G36920, AT2G25820, AT1G78080, AT1G43160, AT5G52020, AT1G22985, AT2G20880, AT3G20840, AT2G46310, AT4G17500, AT1G63030, AT4G13040, AT5G18560, AT4G16750, AT2G44840, AT4G25490, AT5G10510, AT1G44830, AT5G13330, AT1G50680, AT4G32800, AT1G53910, AT3G23240, AT4G23750, AT2G41710, AT4G13620, AT5G13910, AT3G23220, AT5G25810, AT2G47520, AT2G22200, AT5G51190, AT3G54320, AT1G51190, AT1G24590, AT3G11020, AT1G28360, AT3G15210, AT4G37750, AT5G07580, AT5G11190, AT1G25470, AT1G22190, AT1G72360, AT1G75490, AT1G49120, AT5G60120, AT3G23230, AT4G27950, AT3G57600, AT5G47230, AT1G15360, AT3G25890, AT1G28370, AT1G53170, AT5G67190, AT2G28550, AT5G25390, AT5G19790, AT1G77200, AT1G74930, AT1G68550, AT1G33760, AT5G11590, AT3G16280, AT1G03800, AT1G68840, AT1G12980, AT5G07310, AT4G31060, AT2G40220, AT1G01250, AT5G61590, AT1G16060, AT4G06746, AT5G50080, AT5G67180, AT4G17490, AT3G14230, AT5G67010, AT2G38340, AT5G05410, AT2G40350, AT1G13260, AT1G71450
B3	AT1G35540, AT2G30470, AT2G36080, AT5G32460, AT2G24680, AT4G31690, AT1G43950, AT1G19220, AT4G31660, AT4G03170, AT3G25730, AT4G31610, AT5G62000, AT3G11580, AT2G28350, AT3G06220, AT1G34310, AT5G60142, AT1G35240, AT5G60140, AT3G06160, AT4G31640, AT1G30330, AT1G34390, AT1G34410, AT2G46870, AT5G57720, AT3G61830, AT1G20600, AT5G06250, AT1G51120, AT1G19850, AT4G31650, AT2G24650, AT3G17010, AT4G01580, AT3G53310, AT5G20730, AT5G18000, AT5G42700, AT2G33860, AT1G59750, AT4G31620, AT1G25560, AT5G25470, AT5G58280, AT5G25475, AT2G24700, AT1G01030, AT2G35310, AT5G09780, AT4G34400, AT1G50680, AT2G24645, AT2G16210, AT4G00260, AT4G21550, AT4G01500, AT1G49480, AT4G33280, AT1G34170, AT1G16640, AT4G31630, AT1G77850, AT3G18990, AT4G31615, AT2G33720, AT4G30080, AT3G46770, AT4G23980, AT3G26790, AT3G18960, AT1G08985, AT5G18090, AT1G49475, AT1G26680, AT3G19184, AT2G24681, AT3G61970, AT5G60130, AT1G05930, AT3G24650, AT1G68840, AT5G66980, AT5G37020, AT4G32010, AT2G24690, AT4G31680, AT2G24696, AT1G28300, AT5G60450, AT1G35520, AT1G13260, AT2G46530
bHLH-MYC N	AT5G46760, AT4G09820, AT1G32640, AT4G00870, AT2G31280, AT2G16910, AT4G16430, AT4G17880, AT1G10610, AT1G63650, AT2G46510, AT5G41315, AT1G01260, AT4G00480, AT5G46830
bZIP 1	AT3G54620, AT3G19290, AT5G11260, AT2G17770, AT1G77920, AT5G24800, AT3G51960, AT3G44460, AT2G16770, AT1G32150, AT4G35900, AT3G30530, AT2G18160, AT1G06850, AT1G75390, AT2G46270, AT2G21230, AT5G06839, AT1G22070, AT5G06950, AT4G34590, AT2G40620, AT5G60830, AT4G37730, AT3G49760, AT4G36730, AT4G35040, AT2G35530, AT2G41070, AT3G62420, AT2G42380, AT4G34000, AT4G38900, AT5G28770, AT3G56660, AT2G04038, AT4G02640, AT5G42910, AT1G08320, AT3G17609, AT5G07160, AT2G31370, AT5G38800, AT5G08141, AT5G49450, AT1G68880, AT5G15830, AT3G58120, AT2G12940, AT5G44080, AT2G13150, AT3G10800, AT1G42990, AT2G40950, AT5G06960, AT3G56850, AT1G68640, AT1G13600, AT2G22850, AT4G01120, AT1G43700, AT5G10030, AT1G49720, AT5G65210, AT1G06070, AT3G12250, AT2G36270, AT1G03970, AT1G59530

Supplemental Table 4.7 (cont'd)

bZIP 2	AT3G54620, AT3G19290, AT5G11260, AT2G17770, AT5G24800, AT3G51960, AT3G44460, AT2G16770, AT1G32150, AT4G35900, AT3G30530, AT2G18160, AT1G06850, AT1G75390, AT2G46270, AT2G21230, AT5G06839, AT4G38900, AT5G06950, AT2G22850, AT2G40620, AT5G60830, AT4G37730, AT3G49760, AT5G44080, AT4G36730, AT4G35040, AT2G35530, AT2G41070, AT3G62420, AT2G42380, AT4G34000, AT5G07160, AT5G28770, AT3G56660, AT2G04038, AT4G02640, AT5G42910, AT1G08320, AT3G17609, AT2G31370, AT5G38800, AT5G08141, AT5G49450, AT2G12900, AT1G68880, AT5G15830, AT3G58120, AT2G12940, AT4G34590, AT2G13150, AT3G10800, AT1G42990, AT2G40950, AT5G06960, AT3G56850, AT1G68640, AT1G13600, AT1G45249, AT4G01120, AT1G43700, AT5G10030, AT1G49720, AT5G65210, AT1G06070, AT3G12250, AT2G36270, AT1G03970, AT1G59530
bZIP C	AT4G02640, AT5G24800, AT5G28770, AT3G54620
CBFB NFYA	AT3G05690, AT1G17590, AT3G20910, AT1G30500, AT1G72830, AT2G34720, AT3G14020, AT1G54160, AT5G06510, AT5G12840
E2F TDP	AT3G48160, AT5G14960, AT2G36010, AT1G47870, AT5G02470, AT3G01330, AT5G22220, AT5G03415
EIN3	AT5G21120, AT3G20770, AT2G27050, AT5G65100, AT1G73730, AT5G10120
FAR1	AT5G18960, AT1G52520, AT5G28530, AT2G32250, AT4G19990, AT3G07500, AT1G10240, AT1G76320, AT3G59470, AT4G12850, AT4G15090, AT2G43280, AT2G27110, AT4G38180, AT3G06250, AT3G22170, AT1G80010
GAGA	AT2G01930, AT2G35550, AT1G68120, AT1G14685, AT4G38910, AT5G42520, AT2G21240
GATA	AT1G51600, AT5G56860, AT3G24050, AT3G60530, AT3G21175, AT1G08010, AT5G47140, AT4G34680, AT3G45170, AT5G66320, AT2G28340, AT3G50870, AT5G26930, AT4G36240, AT3G16870, AT5G25830, AT3G06740, AT1G08000, AT3G54810, AT4G24470, AT4G17570, AT3G20750, AT2G45050, AT3G51080, AT2G18380, AT4G36620, AT4G16141, AT4G26150, AT4G32890, AT5G49300
HALZ	AT5G65310, AT2G22430, AT3G60390, AT2G01430, AT5G06710, AT2G46680, AT3G61890, AT3G01220, AT2G22800, AT3G01470, AT1G27045, AT4G37790, AT4G17460, AT5G47370, AT4G40060, AT1G70920, AT1G26960, AT5G15150, AT1G69780, AT2G44910, AT4G16780
HLH	AT5G61270, AT5G37800, AT1G01260, AT4G25400, AT5G43650, AT2G43010, AT2G42300, AT3G26744, AT1G05805, AT1G51140, AT4G02590, AT1G59640, AT5G51780, AT1G74500, AT1G66470, AT1G71200, AT5G65320, AT3G56970, AT5G58010, AT4G21330, AT4G16430, AT1G72210, AT1G10610, AT2G31210, AT5G41315, AT2G31215, AT3G56980, AT3G23690, AT4G33880, AT1G10120, AT5G46690, AT5G43175, AT2G46510, AT3G19860, AT5G01310, AT3G23210, AT2G42280, AT3G47710, AT5G08130, AT4G00050, AT4G28790, AT2G31220, AT5G62610, AT3G59060, AT2G41130, AT2G46970, AT1G18400, AT3G25710, AT1G68920, AT4G09820, AT5G51790, AT5G48560, AT3G24140, AT2G22770, AT2G24260, AT1G22490, AT1G06170, AT2G14760, AT1G68810, AT1G26260, AT1G35460, AT5G46760, AT4G36930, AT2G22760, AT5G54680, AT5G65640, AT1G03040, AT3G61950, AT3G56770, AT2G20180, AT1G12860, AT4G30980, AT5G67060, AT2G46810, AT1G26945, AT2G41240, AT3G07340, AT3G50330, AT2G40200, AT1G12540, AT5G57150, AT5G50915, AT4G17880, AT2G28160, AT1G69010, AT2G22750, AT1G27740, AT4G34530, AT1G02340, AT4G14410, AT4G37850, AT1G09530, AT3G62090, AT1G32640, AT4G00870, AT5G10570, AT5G04150, AT3G06120, AT4G00120, AT4G09180, AT3G21330, AT4G36540, AT4G28800, AT5G67110, AT5G56960, AT1G49770, AT5G38860, AT2G16910, AT2G18300, AT4G20970, AT2G43140, AT5G09750, AT1G68240, AT5G53210, AT5G46830, AT1G62975, AT4G29930, AT4G25410, AT4G28811, AT1G25330, AT4G28815, AT1G63650, AT4G38070, AT1G51070, AT4G01460, AT1G73830
Homeobox KN	AT2G27220, AT1G75430, AT1G62990, AT1G75410, AT2G23760, AT4G36870, AT4G32040, AT1G62360, AT2G35940, AT5G02030, AT5G11060, AT1G70510, AT4G32980, AT1G23380, AT4G25530, AT2G16400, AT5G25220, AT2G27990, AT4G34610, AT5G41410, AT1G19700, AT4G08150

Supplemental Table 4.7 (cont'd)

HSF DNA-bind	AT2G27220, AT1G75430, AT1G62990, AT1G75410, AT2G23760, AT4G36870, AT4G32040, AT1G62360, AT2G35940, AT5G02030, AT5G11060, AT1G70510, AT4G32980, AT1G23380, AT4G25530, AT2G16400, AT5G25220, AT2G27990, AT4G34610, AT5G41410, AT1G19700, AT4G08150
K-box	AT5G20240, AT1G31140, AT2G22540, AT5G65060, AT4G22950, AT5G51860, AT3G57230, AT4G24540, AT5G65080, AT3G02310, AT5G60910, AT1G69120, AT2G45650, AT3G57390, AT1G24260, AT1G77080, AT4G18960, AT5G10140, AT3G61120, AT4G11880, AT5G65050, AT5G51870, AT1G26310, AT4G37940, AT5G23260, AT2G22630, AT4G09960, AT5G62165, AT2G14210, AT3G54340, AT5G65070, AT2G45660, AT3G58780, AT5G15800, AT2G03710, AT2G42830, AT1G71692, AT5G13790, AT3G30260
KNOX1	AT4G32040, AT1G70510, AT5G25220, AT5G11060, AT1G62990, AT1G23380, AT1G62360, AT4G08150
KNOX2	AT4G32040, AT1G70510, AT5G25220, AT5G11060, AT1G62990, AT1G23380, AT1G62360, AT4G08150
MADF DNA bdg	AT1G33240, AT3G10030, AT3G24490, AT1G76880, AT1G76890, AT2G44730
MEKHLA	AT1G30490, AT2G34710, AT5G60690, AT4G32880, AT1G52150
MFMR	AT4G01120, AT2G46270, AT2G35530, AT4G36730, AT1G32150
Myb CC LHEQLE	AT3G12730, AT3G13040, AT4G13640, AT4G28610, AT5G29000, AT3G04030, AT3G04450, AT5G06800, AT5G18240, AT5G45580, AT3G24120, AT1G69580, AT2G20400, AT1G79430, AT2G01060
Myb DNA-bind 4	AT1G31310, AT1G33240, AT3G58630, AT1G76880, AT2G44730, AT2G35640, AT1G21200, AT2G33550, AT3G10030, AT3G24490, AT3G24860, AT1G76870, AT2G38250, AT5G63420, AT5G01380, AT5G03680, AT5G05550, AT1G13450, AT5G47660, AT1G54060, AT1G76890, AT3G11100, AT3G14180, AT3G10040, AT5G28300, AT3G54390, AT4G31270, AT3G10000, AT3G25990
Myb DNA-bind 5	AT1G33240, AT3G25990, AT1G76890, AT1G13450
Myb DNA-bind 6	AT3G50060, AT3G24310, AT3G09370, AT4G38620, AT5G40360, AT2G37630, AT4G25560, AT5G65790, AT1G56160, AT3G01530, AT5G12870, AT4G05100, AT4G32730, AT1G79180, AT3G13540, AT1G48000, AT3G01140, AT4G17785, AT4G37780, AT2G31180, AT3G46130, AT5G14340, AT1G72740, AT3G06490, AT3G52250, AT3G27810, AT1G76890, AT4G13480, AT2G25230, AT5G56110, AT4G21440, AT3G28470, AT5G07700, AT5G14750, AT1G66370, AT4G00540, AT1G18570, AT1G74430, AT4G37260, AT3G49690, AT5G07690, AT5G54230, AT4G33450, AT3G11450, AT2G36890, AT3G27785, AT2G13960, AT5G59780, AT2G32460, AT3G55730, AT3G48920, AT1G08810, AT3G13890, AT2G39880, AT3G10113, AT3G12820, AT1G71030, AT2G47190, AT4G34990, AT1G25340, AT4G22680, AT2G47460, AT1G15720, AT5G40350, AT4G28110, AT1G18960, AT5G52600, AT5G45420, AT3G11280, AT1G18710, AT1G74080, AT3G12720, AT5G41020, AT1G26780, AT3G29020, AT5G67300, AT5G10280, AT5G55020, AT5G61420, AT3G53200, AT5G39700, AT3G02940, AT3G09230, AT1G22640, AT5G62470, AT5G57620, AT5G11050, AT3G27920, AT5G40430, AT5G35550, AT1G34670, AT3G61250, AT1G17520, AT3G49850, AT5G17800, AT5G60890, AT5G01200, AT1G74650, AT1G69560, AT3G11440, AT2G26950, AT5G65230, AT3G28910, AT1G66230, AT1G35515, AT1G73410, AT5G58340, AT5G23000, AT5G11510, AT3G47600, AT5G40330, AT4G01680, AT2G16720, AT3G60460, AT5G62320, AT4G09460, AT1G14350, AT5G06100, AT3G23250, AT1G09540, AT1G57560, AT2G38090, AT1G72650, AT2G46410, AT1G18330, AT5G16600, AT5G15310, AT5G04760, AT4G12350, AT2G23290, AT5G05790, AT4G18770, AT5G58850, AT1G16490, AT1G56650, AT1G58220, AT5G02320, AT5G06110, AT3G30210, AT1G17950, AT1G66380, AT5G16770, AT3G08500, AT5G49330, AT5G49620, AT1G06180, AT1G68320, AT5G58900, AT4G26930, AT3G18100, AT5G52260, AT5G26660, AT1G63910, AT2G02820, AT1G09770, AT2G26960, AT5G67580, AT3G62610, AT1G66390

Supplemental Table 4.7 (cont'd)

Myb DNA-binding	AT3G50060, AT2G25180, AT3G24310, AT3G09370, AT4G38620, AT5G40360, AT2G37630, AT3G57980, AT5G05790, AT2G44430, AT5G65790, AT1G56160, AT3G01530, AT5G12870, AT4G05100, AT1G01380, AT4G32730, AT1G79180, AT5G02840, AT3G13540, AT3G25790, AT4G18020, AT3G10590, AT4G09450, AT5G59570, AT3G04450, AT1G48000, AT3G01140, AT4G17785, AT4G37780, AT1G66230, AT2G31180, AT3G46130, AT5G14340, AT4G13640, AT1G72740, AT2G18328, AT3G52250, AT1G70000, AT1G06910, AT3G27810, AT4G13480, AT2G40970, AT4G04580, AT2G25230, AT1G66380, AT5G39700, AT5G56110, AT4G21440, AT3G28470, AT5G07700, AT5G08520, AT4G36570, AT1G66370, AT4G25560, AT4G00540, AT1G18570, AT1G22640, AT4G37260, AT3G49690, AT5G07690, AT5G54230, AT4G33450, AT3G11450, AT2G36890, AT3G27785, AT2G13960, AT5G59780, AT3G08500, AT2G32460, AT5G26660, AT1G17460, AT5G58080, AT1G69580, AT3G55730, AT5G18240, AT5G53200, AT3G12730, AT3G48920, AT4G16110, AT1G08810, AT3G13890, AT2G39880, AT3G10113, AT3G12820, AT1G71030, AT5G06100, AT2G47190, AT4G34990, AT1G25340, AT3G10760, AT4G22680, AT2G47460, AT4G37180, AT1G15720, AT5G40350, AT1G09710, AT3G53790, AT5G67300, AT5G41020, AT4G28110, AT1G18960, AT5G52600, AT4G39250, AT3G11280, AT2G06020, AT1G79430, AT2G01060, AT1G74080, AT3G12720, AT5G44190, AT4G31920, AT5G16770, AT1G26780, AT3G29020, AT2G42660, AT5G61620, AT3G10580, AT5G10280, AT5G55020, AT1G18330, AT4G17695, AT3G53200, AT2G38300, AT3G02940, AT3G09230, AT5G62470, AT5G57620, AT5G11050, AT3G60460, AT5G23650, AT3G46640, AT3G27920, AT2G30432, AT5G40430, AT5G35550, AT5G45420, AT3G62670, AT1G34670, AT3G61250, AT1G17520, AT1G74430, AT3G49850, AT5G17800, AT5G60890, AT4G28610, AT5G29000, AT1G18710, AT5G01200, AT1G74650, AT1G13300, AT1G68670, AT1G69560, AT3G11440, AT2G26950, AT2G40260, AT4G39160, AT3G28910, AT1G49950, AT5G05090, AT1G35515, AT1G73410, AT3G13040, AT5G37260, AT5G23000, AT1G49010, AT3G47600, AT5G40330, AT3G04030, AT4G01680, AT2G16720, AT2G30420, AT5G62320, AT2G30424, AT3G24120, AT4G09460, AT3G21430, AT4G01280, AT5G07210, AT5G47390, AT1G14350, AT1G14600, AT3G23250, AT2G20570, AT1G09540, AT1G57560, AT5G11510, AT1G19000, AT2G38090, AT1G72650, AT2G46410, AT2G03500, AT5G16600, AT5G15310, AT3G09600, AT5G17300, AT5G04760, AT5G65230, AT5G56840, AT1G01520, AT3G16350, AT4G12350, AT2G01760, AT2G23290, AT5G14750, AT5G42630, AT1G67710, AT4G18770, AT1G32240, AT5G58340, AT5G58850, AT1G74840, AT1G16490, AT3G60110, AT1G56650, AT1G58220, AT5G02320, AT5G06110, AT3G30210, AT2G20400, AT1G17950, AT1G25550, AT5G16560, AT3G16857, AT1G49560, AT3G06490, AT5G49330, AT5G49620, AT5G06800, AT2G36960, AT1G06180, AT1G68320, AT1G01060, AT5G58900, AT5G61420, AT4G26930, AT3G18100, AT5G52260, AT4G01060, AT1G63910, AT2G02820, AT1G09770, AT5G52660, AT2G26960, AT2G46830, AT5G45580, AT5G67580, AT3G62610, AT2G02060, AT1G66390, AT2G42150
NAM	AT3G49530, AT5G64530, AT5G39690, AT5G63790, AT3G61910, AT3G15510, AT5G62380, AT5G04400, AT1G79580, AT1G60350, AT4G36160, AT5G56620, AT3G01600, AT2G46770, AT3G55210, AT1G60280, AT5G46590, AT5G13180, AT2G43000, AT2G33480, AT1G03490, AT5G09330, AT2G27300, AT1G19040, AT4G01550, AT1G54330, AT3G10500, AT5G07680, AT1G71930, AT3G56530, AT1G02250, AT5G66300, AT4G27410, AT5G50820, AT1G76420, AT3G04070, AT4G01520, AT1G32510, AT4G17980, AT3G56520, AT3G44350, AT1G60380, AT5G39820, AT1G25580, AT5G22290, AT3G15170, AT1G60300, AT5G24590, AT3G04060, AT5G18270, AT5G64060, AT5G14000, AT3G18400, AT5G08790, AT1G64105, AT1G02230, AT3G29035, AT3G04410, AT5G41090, AT3G44290, AT1G65910, AT4G10350, AT5G39610, AT3G10480, AT4G29230, AT4G28500, AT1G12260, AT1G56010, AT1G52890, AT2G24430, AT1G28470, AT1G01010, AT1G69490, AT1G33280, AT5G61430, AT1G02220, AT2G17040, AT5G14490, AT3G12910, AT1G60240, AT3G15500, AT1G77450, AT1G32770, AT1G32870, AT5G18037, AT3G10490, AT3G17730, AT4G28530, AT3G04430, AT3G03200, AT1G01720, AT1G02210, AT2G02450, AT5G22380, AT1G34190, AT5G18300, AT4G01540, AT1G52880, AT4G35580, AT3G56560, AT3G12977, AT3G04420, AT1G33060, AT1G61110, AT1G26870, AT2G18060, AT5G04410, AT5G17260, AT5G53950, AT1G60340, AT1G62700, AT1G34180

Supplemental Table 4.7 (cont'd)

Plant zn clust	AT2G23320, AT5G28650, AT3G04670, AT2G30590, AT2G24570, AT4G31550, AT4G24240
SBP	AT1G76580, AT1G69170, AT5G50670, AT1G02065, AT3G60030, AT2G42200, AT1G53160, AT5G50570, AT1G20980, AT5G18830, AT3G57920, AT2G33810, AT2G47070, AT5G43270, AT1G27370, AT1G27360, AT3G15270
SRF-TF	AT1G31140, AT4G22950, AT5G15800, AT5G60910, AT3G66656, AT1G69120, AT2G26320, AT5G39750, AT5G60440, AT3G04100, AT1G65360, AT5G49420, AT1G65300, AT2G24840, AT1G26310, AT4G37940, AT5G27130, AT5G62165, AT2G14210, AT1G47760, AT3G05860, AT1G60040, AT5G27580, AT5G20240, AT2G22540, AT4G36590, AT1G65330, AT5G49490, AT5G58890, AT4G02235, AT2G34440, AT2G45650, AT5G26630, AT2G03060, AT3G57390, AT1G77080, AT5G10140, AT1G77950, AT1G72350, AT1G31630, AT2G22630, AT5G51870, AT5G48670, AT3G54340, AT1G28450, AT5G27810, AT2G03710, AT3G61120, AT1G22130, AT1G22590, AT5G51860, AT3G57230, AT4G24540, AT5G65080, AT1G46408, AT5G40120, AT3G02310, AT5G41200, AT5G38740, AT5G39810, AT5G65060, AT5G26650, AT5G27944, AT5G38620, AT5G27090, AT5G04640, AT1G60920, AT5G06500, AT4G09960, AT5G40220, AT5G65070, AT5G27070, AT1G01530, AT2G45660, AT5G26950, AT3G18650, AT1G77980, AT1G17310, AT1G48150, AT5G26880, AT2G40210, AT5G37415, AT5G65050, AT1G28460, AT1G59810, AT5G27960, AT1G69540, AT1G60880, AT5G55690, AT1G24260, AT4G18960, AT2G42830, AT2G28700, AT4G11880, AT1G29962, AT5G23260, AT5G27050, AT4G11250, AT1G18750, AT3G58780, AT5G26580, AT1G71692, AT5G65330, AT5G13790, AT3G30260, AT1G31640
SWIM	AT5G18960, AT5G28530, AT2G32250, AT4G19990, AT1G10240, AT1G76320, AT4G15090, AT2G27110, AT4G38180, AT3G06250, AT1G80010
TCP	AT1G53230, AT5G51910, AT1G35560, AT5G08070, AT1G30210, AT1G72010, AT3G45150, AT2G37000, AT5G41030, AT3G27010, AT3G02150, AT1G58100, AT1G68800, AT5G23280, AT3G47620, AT1G67260, AT2G31070, AT1G69690, AT5G08330, AT3G18550, AT4G18390, AT3G15030, AT2G45680, AT5G60970
WRKY	AT5G45260, AT5G49520, AT1G30650, AT1G29280, AT1G80840, AT2G38470, AT2G30250, AT1G64000, AT2G34830, AT4G04450, AT3G58710, AT4G23550, AT2G47260, AT2G46130, AT5G46350, AT2G04880, AT5G56270, AT4G26440, AT2G37260, AT1G13960, AT4G12020, AT1G69310, AT2G40740, AT5G24110, AT2G30590, AT4G01720, AT5G01900, AT5G64810, AT1G29860, AT5G52830, AT1G80590, AT3G04670, AT5G07100, AT1G66550, AT4G31550, AT2G46400, AT1G62300, AT2G23320, AT2G03340, AT2G21900, AT2G44745, AT3G56400, AT1G18860, AT5G22570, AT4G39410, AT3G62340, AT5G41570, AT5G13080, AT3G01080, AT1G69810, AT4G30935, AT4G22070, AT4G11070, AT4G01250, AT1G68150, AT4G24240, AT1G66600, AT4G18170, AT5G28650, AT4G31800, AT2G24570, AT2G25000, AT1G55600, AT5G45050, AT4G23810, AT5G43290, AT4G26640, AT5G26170, AT2G40750, AT3G01970, AT5G15130, AT1G66560
YABBY	AT4G00180, AT2G45190, AT1G69180, AT1G23420, AT2G26580, AT1G08465
zf-B box	AT5G57660, AT1G25440, AT4G10240, AT1G28050, AT5G15840, AT4G38960, AT2G31380, AT1G78600, AT1G06040, AT3G07650, AT4G39070, AT1G75540, AT2G47890, AT5G15850, AT4G15250, AT2G21320, AT2G24790, AT1G73870, AT2G33500, AT1G68520, AT5G48250, AT5G24930, AT3G02380
zf-C2H2	AT3G10470, AT1G14580, AT1G02030, AT2G41835, AT3G48430, AT4G06634, AT5G66730, AT2G45120, AT5G04390, AT1G26590, AT2G02070, AT3G45260, AT5G61470, AT5G03150, AT1G72050, AT3G57670, AT2G17180, AT1G55110, AT3G60580, AT5G56200, AT1G68130, AT2G02080, AT5G14140
zf-C2H2 4	AT5G61470, AT3G48430, AT3G20880, AT3G60580, AT1G08290, AT1G02030, AT1G51220, AT1G13290, AT5G14140, AT1G72050, AT1G34790, AT2G45120, AT1G26590, AT4G06634

Supplemental Table 4.7 (cont'd)

zf-C2H2 6	AT3G10470, AT4G16610, AT5G15480, AT5G59820, AT4G17810, AT5G67450, AT3G23130, AT1G02030, AT5G01860, AT3G46080, AT3G53600, AT1G27730, AT2G28200, AT3G49930, AT2G42410, AT2G45120, AT1G26590, AT4G35280, AT3G19580, AT5G61470, AT3G46070, AT2G37740, AT3G46090, AT2G17180, AT5G03510, AT3G29340, AT3G60580, AT1G26610, AT2G37430, AT5G04390, AT5G43170, AT5G56200, AT2G28710, AT5G04340
zf-C2H2 jaz	AT5G61470, AT1G27730, AT1G74250, AT2G28200, AT5G56200, AT5G26749, AT2G45120, AT1G02030, AT2G26940, AT2G37430, AT1G26590, AT2G17180, AT3G60580
zf-CCCH	AT5G06770, AT3G12130, AT5G49200, AT3G12680, AT5G44260, AT3G19360, AT1G29560, AT1G48195, AT5G06420, AT5G16540, AT1G32360, AT2G28450, AT5G58620, AT3G06410, AT1G04990, AT2G47850, AT3G21810, AT1G29600, AT3G02830, AT2G32930, AT5G18550, AT3G08505, AT1G10320, AT1G29570, AT1G68200, AT1G01350, AT5G63260, AT2G35430, AT1G66810, AT1G75340, AT3G44785, AT3G48440, AT2G25900, AT1G03790
zf-Dof	AT2G37590, AT1G51700, AT2G46590, AT5G60850, AT1G47655, AT3G47500, AT3G61850, AT1G64620, AT1G29160, AT3G21270, AT1G28310, AT5G65590, AT1G26790, AT4G38000, AT5G02460, AT2G34140, AT2G28510, AT4G21050, AT5G66940, AT3G45610, AT1G07640, AT4G00940, AT4G21030, AT4G24060, AT3G52440, AT3G50410, AT3G55370, AT5G62940, AT5G39660, AT4G21080, AT1G69570, AT1G21340, AT5G60200, AT4G21040, AT5G62430, AT2G28810
ZF-HD dimer	AT3G50890, AT5G60480, AT5G15210, AT2G02540, AT1G69600, AT1G14687, AT4G24660, AT2G18350, AT3G28920, AT1G18835, AT5G42780, AT1G74660, AT3G28917, AT1G75240, AT5G65410, AT1G14440, AT5G39760
zf-met	AT5G61470, AT1G27730, AT1G74250, AT1G02030

Supplemental Table 4.8 Best fit parameters of ODE models of the evolution of TF expression above or below the ancestral state

Number of Parameters	Data	x	y	z	w	mu[1]	Sigma[2]	-2 log(L)[3]
1	Control	0.089	0.089	0.089	0.089	0.002	0.002	-90.6
1	LightDev	0.095	0.095	0.095	0.095	0.004	0.003	-80.2
1	StressDiff	0.081	0.081	0.081	0.081	0.001	0.001	-100.7
1	Treatment	0.090	0.090	0.090	0.090	0.001	0.001	-94.0
2	Control	0.113	0.113	0.067	0.067	0.001	0.001	-98.0
2	LightDev	0.148	0.148	0.056	0.056	0.001	0.002	-125.0
2	StressDiff	0.074	0.074	0.087	0.087	0.001	0.001	-101.3
2	Treatment	0.121	0.121	0.064	0.064	0.001	0.001	-103.0

1. Mean of error distribution
2. Standard deviation of error distribution
3. -2 log-likelihood of model fit

Supplemental Table 4.9 Best fit parameters of ODE models of partitioning of ancestral states between duplicate TFs

Number of Parameters	Data	x	y	z	w	mu[1]	sigma[2]	-2 log L[3]
1	Control	0.537	0.537	0.537	0.537	0.010	0.010	-75.5
1	LightDev	0.5298	0.5298	0.5298	0.5298	0.010	0.010	-75.3
1	StressDiff	0.471	0.471	0.471	0.471	0.008	0.008	-81.4
1	Treatment	0.054	0.054	0.054	0.05	0.010	0.010	-77.5
1	DapSeq	1.756	1.756	1.756	1.756	0.041	0.048	-38.8
2	Control	0.957	0.957	0.0779	0.0779	0.001	0.001	-140.3
2	LightDev	0.978	0.978	0.08	0.08	0.001	0.001	-139.4
2	StressDiff	0.747	0.747	0.0669	0.0669	0.001	0.001	-140.2
2	Treatment	1.368	1.368	0.0749	0.0749	1.70E-05	0.001	-137.7
2	DapSeq	40.43	40.43	0.222	0.222	0.031	0.031	-48.8
4	DapSeq	10.12	4.1	0.2063	0.4175	0.003	0.004	-97.8

1. Mean of error distribution
2. Standard deviation of error distribution
3. -2 log-likelihood of model fit

Supplemental Table 4.10 The importance of all features used in the classification of individual duplicate genes

Feature	Genome[1]	Kinases[1]	TFs[1]
Maximum Percent Identity (Paralog)	171.6 (1)	57.9 (1)	29.4 (2)
Sequence Conservation (Viridiplantae)	109.3 (2)	27.2 (2)	31.8 (1)
Gene Family Size (OrthoMCL)	81.6 (3)	2.0 (19)	27.7 (3)
Protein Length (in Amino Acids)	52.5 (4)	14.2 (3)	10.5 (11)
Expression Breadth (AtGenExpress)	46.2 (5)	9.4 (10)	11.2 (8)
Expression MAD/Median (AtGenExpress)	41.0 (6)	5.4 (11)	11.8 (5)
Expression Mean (LightDev Data)	40.8 (7)	10.4 (7)	10.7 (9)
Expression Breadth (RNASeq)	40.0 (8)	3.3 (15)	11.5 (6)
Expression Mean (Control Data)	39.2 (9)	10.7 (6)	10.7 (10)
Expression Median (AtGenExpress)	37.5 (10)	10.2 (8)	10.2 (12)
Expression Mean (Stress Data)	37.0 (11)	11.5 (4)	12.0 (4)
Expression Mean (AtGenExpress)	36.9 (12)	11.3 (5)	11.3 (7)
Expression Median (RNASeq)	34.8 (13)	4.7 (12)	4.9 (16)
Sequence Conservation (Metazoa)	34.5 (14)	3.6 (13)	4.4 (18)
Nucleotide Diversity (Pi)	32.4 (15)	9.7 (9)	6.2 (14)
Expression Mean (Diff Data)	31.6 (16)	1.7 (2)	7.9 (13)
Sequence Conservation (Fungi)	30.6 (17)	2.4 (17)	4.6 (17)
Number of Protein Domains	28.4 (18)	2.3 (18)	5.6 (15)
Expression Mean (RNASeq)	18.6 (19)	3.0 (16)	2.9 (19)
Expression Maximum (RNASeq)	12.9 (20)	3.3 (14)	0.4 (20)

1: The importance of the feature as defined by the mean decrease in accuracy of the classification when the feature is removed. Features are ordered according to the rank of their importance in the whole genome model and the rank of each value for each model is indicated by ().

CHAPTER 5: CONCLUSION

In the preceding chapters, I have presented the application of mathematical modeling techniques to my research, including the characterization of cyclic expression in two species and the evolution of TF function and regulation. Here, I will further discuss the results of these models as well as future directions for my research specifically and the application of modeling to the understanding of gene expression in general.

PREDICTING CYCLIC EXPRESSION PATTERNS USING *CIS*-REGULATORY ELEMENTS

In the work presented in Chapter 2, we identified genes that were cyclically expressed under diel conditions in the algae *C. reinhardtii* and clustered them according their phase of expression. Using these phase clusters, we found pCREs enriched in the promoters of genes sharing the same phase of expression and trained a SVM model of diel expression using these pCREs as features. The resulting model was able to predict the expression phase of diel genes better than both random guessing and naive classifiers, but performed worse than previous SVM models for predicting stress response in *A. thaliana* (Zou et al., 2011). However, we were able to improve the performance of our classifier for subsets of cyclic genes by subdividing phase-clusters according to enriched gene functions. This improvement was not unexpected, given the strict association between annotated gene function and phase of expression observed in our study, but still indicates that the identified pCREs are important for regulating a subset of diel expressed genes.

Continuing to look at cyclic expression, Chapter 3 describes a project where we changed the system to cell-cycle expression in *S. cerevisiae* in order to take advantage of the extensive expression and regulatory data available in this species. Here, we found that classifiers built using TF-target interactions derived from ChIP-Chip, TF Deletion, and PWM data performed

better at predicting the phase of cell-cycle expression than the pCRE model of *C. reinhardtii* diel expression. This improved performance is in part dependent on the near complete coverage of TFs in *S. cerevisiae* by each data set, as reducing coverage of the best performing data sets to less than half of the total *S. cerevisiae* TFs resulted in models with similar performance to those in *C. reinhardtii*. Conversely, performance was improved by using interactions between TFs that were part of regulatory FFLs and by combining features from the ChIP-Chip and Deletion data sets. Subsequent importance and network analysis of features revealed that both canonical cell-cycle regulators and at least two modules of TFs which lack evidence of being cell-cycle regulators are amongst the best predictors of cell-cycle expression.

Nevertheless, even in our final model of *S. cerevisiae* cell-cycle regulation, there is room to improve classification and learn more about the regulatory mechanisms controlling expression. Using the optimal scoring threshold, our best cell-cycle model had an ~5% false positive rate across the different phase classes, but recovered only half of the known cell-cycle genes in each class. Why are these genes improperly classified? What aspects of cell-cycle regulation are missing from our model? In the case of cell-cycle expression, it is known that cyclin-dependent kinases play a key role in controlling cell-cycle progression both through direct regulation of cell-cycle proteins and indirectly by modulating TF activity (Csikász-Nagy et al., 2009). In particular, phosphorylation controls the activity of Swi6 (Sidorova et al., 1995; Lomberk et al., 2006; Shimada et al., 2009), a known regulator of cell-cycle initiation (Ho et al., 1999) and one of the most important features in our model of cell-cycle expression. Swi6 itself participates in chromatin remodeling (Grewal and Elgin, 2002; Haldar et al., 2011), thus cyclin-dependent kinases may further, indirectly regulate TF activity by affecting DNA accessibility. Likewise, posttranscriptional regulation is of transcription is thought to play a role in

maintenance of diel and circadian cycles (Kojima et al., 2011; Romanowski and Yanovsky, 2015). Altering the expression of RNA-binding proteins in *C. reinhardtii* can not only lead to the loss of cyclic expression, but disrupt the normal phase of cycling genes (Iliev et al. 2006), indicating the post-transcriptional regulation exerts fine-grain control of the timing of circadian and diel cycles. These additional layers of regulation, protein activation, DNA accessibility and post-transcriptional regulation, likely hold the key to improving models of cyclic expression beyond what is able to be done with CREs and pCREs alone.

DUPLICATION AND EVOLUTION OF TRANSCRIPTION FACTORS

While our previous models focused on the relationship between existing TFs and their target genes, in model described in Chapter 4 we looked at the retention of duplicate genes pairs following WGD events and specifically inferred the changes in expression and regulation of TFs post-duplication. We found that TFs duplicates are retained more often than expected compared to genes with other general functions (i.e. kinases, transporter activity, defense response), owing in part to the fact that sequence divergence between duplicate TFs is on average greater than that of other duplicate genes. Looking further into the relationship between duplicate TFs, we found that the loss of an ancestral expression pattern or CREs from one duplicate copy but not the other occurred more frequently than expected by random chance. Furthermore, loss of ancestral expression and CREs occurred asymmetrically, such that duplicated TF pairs could be divided into distinct “ancestral” copy, which retains almost all ancestral expression and regulatory sites, and “non-ancestral” copy, which losses almost all ancestral expression and cis-regulatory sites, but gains novel cis-regulatory sites instead. Furthermore, the partitioning of ancestral states is not random as loss of ancestral expression in the first duplicate copy occurs at approximately an order of magnitude faster than in the second duplicate copy, and the loss of ancestral regulation

occurs two orders of magnitude faster. We theorize that the preference for partitioning retained duplicate TFs in ancestral and non-ancestral copies is due to the neofunctionalization of the non-ancestral copy. This hypothesis is supported by examples where there is experimental evidence supporting neo-functionalization of the non-ancestral copy of a TF duplicate pair.

However, expression and regulation are, at best, only a proxy for biological function. Our model of TF evolution is, therefore, restricted by the fact that we cannot characterize gain or loss of specific functions in the same way we can identify changes in expression or *cis*-regulatory sites. The appropriate way to define biological function has been the subject of some controversy (ENCODE Project Consortium, 2012; Doolittle, 2013; Graur et al., 2013; Brunet and Doolittle, 2014; Kellis et al., 2014a, Kellis et al., 2014b), and while a definition of function relying of biochemical activity is useful for the purpose of broad categorization (e.g. transcription factor, kinase, etc.), when considering gene evolution and retention, the definition of biological function requires evidence of selection as a criterion. Assessing selection on large scale is difficult, even under laboratory conditions (Winzeler et al., 1999; Tong et al., 2001), so an extensive catalog of selective function would be challenging to produce, particularly for organisms with comparably larger genomes and longer life-cycles. Thus, modeling the evolution of functions in the same way we modeled the evolution of expression and regulation may not be possible, but a different type of biological model might offer us an alternative approach. Recent studies have shown that essential (Lloyd et al., 2015) and functional (Tsai et al., 2017) regions of the genome can be classified by their molecular, genetic, and evolutionary features, including expression, conservation, and interactions. Classification is done by scoring regions/genes based on their features and comparing that score to the distribution of scores for genes that are known to be essential, genes with known functions, and putatively non-functional regions such as

pseudogenes. In this way, the likelihood that a gene is essential, functional, or non-functional can be quantified. Therefore, the evolution of the functions of retained duplicates may be explored in a probabilistic manner by comparing the so called "functional-likelihood" score of non-ancestral duplicates across different groups of genes as well as to the functional likelihood of ancestral duplicate copies.

FUTURE PROSPECTS FOR MODELING GENE EXPRESSION

Modeling gene expression, particularly complex patterns sensitive to variations in time, location or condition, is poised to benefit greatly from the continued improvement of sequencing technology and computational methods. Independent of any particular sequencing approach, the declining cost of sequencing DNA (Wetterstrand 2016) means that quantifying expression under broader sets of conditions or with greater detail have become progressively more feasible. For circadian and diel expression in particular, the increase in sampling resolution has generally lead to the identification of more genes having cyclic patterns of expression (Harmer et al., 2000; Edwards et al., 2006; Michael et al., 2008; Panchy et al., 2014; Zones et al., 2015). Long read sequencing technologies, such as PacBio, also hold the promise of improving how we quantify expression by removing the need to infer expression from short-reads that represent a fragment of the actual transcript. Currently, full-length transcript sequencing data from PacBio have been used to reassemble and re-annotate the genome of *Triticum aestivum* (Liu et al., 2017) as well as profile the transcriptomes of *Schizosaccharomyces pombe* (Kuang et al., 2017) and *Oryctolagus cuniculus* (Chen et al., 2017; Liu et al., 2017). However, short-read Illumina sequences are needed in conjunction with PacBio reads due to the high per-base error rate of this technology (Chen et al., 2017; Liu et al., 2017). As PacBio continues to improve and/or other technologies become available that can offer the same long-read sequencing with fewer errors we can expect

to see transcriptome annotation and transcription quantification done using full-length transcript sequences alone. In addition to less costly, higher quality quantification of expression, advances in computing offer new opportunities for expression modeling. Specifically, so called “deep learning” approaches involving neural networks have provided solutions to problems where learning approaches previously had failed (Silver et al., 2016; Litjens et al., 2017). Yet, like other machine learning methods, the models created by deep learning are a “black box” when it comes to deriving biological significance (Albrecht et al., 2017). Though these new technologies offer great potential for both generating and analyzing expression data, it will be up to enterprising modelers to find way to apply these new approaches to answer biological questions.

REFERENCES

REFERENCES

Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. Nat Rev Genet **16**: 197–212

Albert R, Othmer HG (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. J Theor Biol **223**: 1–18

Albrecht T, Slabaugh G, Alonso E, Al-Arif SMMR (2017) Deep learning for single-molecule science. Nanotechnology **28**: 423001

Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol **33**: 831–838

Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet **8**: 450–461

Alon U (2007) An Introduction to Systems Biology. Chapman & Hall/CRC, London

Alvarez-Ponce D, Fare M (2012) Evolutionary rate and duplicability in the Arabidopsis thaliana protein-protein interaction network. Genome Biol Evol **4**: 1247–1263

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol **11**: R106

Archambault V, Li CX, Tackett AJ, Wasch R, Chait BT, Rout MP, Cross FR (2003) Genetic and biochemical evaluation of the importance of Cdc6 in regulating mitotic exit. Mol Biol Cell **14**: 4592–4604

Arimura G, Kopke S, Kunert M, Volpe V, David A, Brand P, Dabrowska P, Maffei ME, Boland W (2008) Effects of feeding Spodoptera littoralis on lima bean leaves: IV. Diurnal and nocturnal damage differentially initiate plant volatile emission. Plant Physiol **146**: 965–973

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25**: 25–29

Ay A, Arnosti DN (2011) Mathematical modeling of gene expression: a guide for the perplexed biologist. Crit Rev Biochem Mol Biol **46**: 137–151

Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, et al (2009) Diversity and complexity in DNA recognition by transcription factors. Science **324**: 1720–1723

Bahler J (2005) Cell-cycle control of gene expression in budding and fission yeast. Annu Rev Genet **39**: 69–94

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res **37**: W202--8

Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res **34**: W369--73

Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol **9**: e1003326

Baker CR, Hanson-Smith V, Johnson AD (2013) Following gene duplication, paralog interference constrains transcriptional circuit evolution. Science **342**: 104–108

Baldwin IT, Meldau S (2013) Just in time: circadian defense patterns and the optimal defense hypothesis. Plant Signal Behav **8**: e24410

Bansal M, Della Gatta G, di Bernardo D (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics **22**: 815–822

Bean JM, Siggia ED, Cross FR (2005) High functional overlap between MluI cell-cycle box binding factor and Swi4/6 cell-cycle box binding factor in the G1/S transcriptional program in Saccharomyces cerevisiae. Genetics **171**: 49–61

Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. Cell **117**: 185–198

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B-Methodological **57**: 289–300

Benveniste D, Sonntag H-J, Sanguinetti G, Sproul D (2014) Transcription factor binding predicts histone modifications in human cell lines. Proc Natl Acad Sci U S A **111**: 13367–13372

Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc **4**: 393–411

Bergholdt R, Brorsson C, Lage K, Nielsen JH, Brunak S, Pociot F (2009) Expression profiling of human genetic and protein interaction networks in type 1 diabetes. PLoS One **4**: e6250

Bertoli C, Skotheim JM, de Bruin RAM (2013) Control of cell cycle transcription during G1 and S phases. Nat Rev Mol Cell Biol **14**: 518–528

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R (2005) Transcriptional regulation by the numbers: models. Curr Opin Genet Dev **15**: 116–124

Birchler JA, Veitia RA (2007) The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell **19**: 395–402

Birchler JA, Veitia RA (2010) The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. New Phytol **186**: 54–62

Bisova K, Krylov DM, Umen JG (2005) Genome-wide annotation and expression profiling of cell cycle regulatory genes in Chlamydomonas reinhardtii. Plant Physiol **137**: 475–491

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16**: 1679–1691

Blasing OE, Gibon Y, Gunther M, Hohne M, Morcuende R, Osuna D, Thimm O, Usadel B, Scheible WR, Stitt M (2005) Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. Plant Cell **17**: 3257–3281

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422**: 433–438

Breeden L (2003) Periodic transcription: a cycle within a cycle. Curr Biol **13**: 31–38

Bruce VG (1970) The biological clock in Chlamydomonas reinhardtii. J Protozool **17**: 328–334

Brunet TDP, Doolittle WF (2014) Getting “function” right. Proc Natl Acad Sci U S A **111**: E3365

Buck MJ, Lieb JD (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics **83**: 349–360

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. **11**: 94

Bulyk ML (2007) Protein binding microarrays for the characterization of DNA-protein interactions. Adv Biochem Eng Biotechnol **104**: 65–85

Bustin SA (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol **25**: 169–193

Byrne TE, Wells MR, Johnson CH (1992) Circadian rhythms of chemotaxis to ammonium and of methylammonium uptake in chlamydomonas. Plant Physiol **98**: 879–886

Carretero-Paulet L, Fares MA (2012) Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Mol Biol Evol **29**: 3541–3551

Cavalier-Smith T (1974) Basal body and flagellar development during the vegetative cell cycle and the sexual cycle of Chlamydomonas reinhardii. J Cell Sci **16**: 529–556

Chen K, Durand D, Farach-Colton M (2000) NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol **7**: 429–447

Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004) Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell **15**: 3841–3862

Chen S-Y, Deng F, Jia X, Li C, Lai S-J (2017) A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Sci Rep **7**: 7648

Chen YB, Dominic B, Mellon MT, Zehr JP (1998) Circadian rhythm of nitrogenase gene expression in the diazotrophic filamentous nonheterocystous cyanobacterium Trichodesmium sp. strain IMS 101. J Bacteriol **180**: 3598–3605

Chen Y, Li Y, Narayan R, Subramanian A, Xie X (2016) Gene expression inference with deep learning. Bioinformatics **32**: 1832–1839

Chen Y, Elenee Argentinis JD, Weber G (2016) IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. Clin Ther **38**: 688–701

Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biol **12**: R15

Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG (2009) Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. PLoS Comput Biol **5**: e1000417

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak M, Gaffney DJ, Elo LL, Zhang X, et al (2016) A survey of best practices for RNA-seq data analysis. Genome Biol **17**: 13

Cooper-Knock J, Kirby J, Ferraiuolo L, Heath PR, Rattray M, Shaw PJ (2012) Gene expression profiling in human neurodegenerative disease. Nat Rev Neurol **8**: 518–530

Corellou F, Schwartz C, Motta JP, Djouani-Tahri el B, Sanchez F, Bouget FY (2009) Clocks in the green lineage: comparative functional analysis of the circadian architecture of the picoeukaryote ostreococcus. Plant Cell **21**: 3436–3449

Csikász-Nagy A, Kapuy O, Tóth A, Pál C, Jensen LJ, Uhlmann F, Tyson JJ, Novák B (2009) Cell cycle regulation by feed-forward loops coupling transcription and phosphorylation. Mol Syst Biol **5**: 236

Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, Cheung VG, Kraus WL, Lis JT, Siepel A (2015) Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods **12**: 433–438

Das MK, Dai H-K (2007) A survey of DNA motif finding algorithms. BMC Bioinformatics **8 Suppl 7**: S21

Davies JP, Grossman AR (1994) Sequences controlling transcription of the Chlamydomonas reinhardtii beta 2-tubulin gene after deflagellation and during the cell cycle. Mol Cell Biol **14**: 5165–5174

Davis S, Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics **23**: 1846–1847

De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. Trends Ecol Evol **20**: 591–597

de Boer CG, Hughes TR (2012) YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic Acids Res **40**: D169--79

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol **3**: e314

Doherty CJ, Kay SA (2010) Circadian control of global gene expression patterns. Annu Rev Genet **44**: 419–444

Dong S, Adams KL (2011) Differential contributions to the transcriptome of duplicated genes in response to abiotic stresses in natural and synthetic polyploids. New Phytol **190**: 1045–1057

Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R, Birney E, et al (2012) Modeling gene expression using chromatin features in various cellular contexts. Genome Biol **13**: R53

Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci U S A **110**: 5294–5300

Edwards KD, Anderson PE, Hall A, Salathia NS, Locke JCW, Lynn JR, Straume M, Smith JQ, Millar AJ (2006) FLOWERING LOCUS C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock. Plant Cell **18**: 639–650

Eldar A, Dorfman R, Weiss D, Ashe H, Shilo B-Z, Barkai N (2002) Robustness of the BMP morphogen gradient in Drosophila embryonic patterning. Nature **419**: 304–308

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature **489**: 57–74

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol **5**: e8

Fang SC, de los Reyes C, Umen JG (2006) Cell size checkpoint control by the retinoblastoma tumor suppressor pathway. PLoS Genet **2**: e167

Farre EM (2012) The regulation of plant growth by the circadian clock. Plant Biol **14**: 401–410

Fay M (2010) Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data. R J **2**: 53–58

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. . Distrib. by author.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics **5**: 164–166

Femino AM, Fay FS, Fogarty K, Singer RH (1998) Visualization of single RNA transcripts in situ. Science (80-) **280**: 585–590

Feng J, Liu T, Qin B, Zhang Y, Liu XS (2012) Identifying ChIP-seq enrichment using MACS. Nat Protoc **7**: 1728–1740

Fernyhough P (2001) Quantification of mRNA levels using northern blotting. Methods Mol Biol **169**: 53–63

Filichkin SA, Breton G, Priest HD, Dharmawardhana P, Jaiswal P, Fox SE, Michael TP, Chory J, Kay SA, Mockler TC (2011) Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. PLoS One **6**: e16907

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res **44**: D279--85

Fonken LK, Nelson RJ (2014) The Effects of Light at Night on Circadian Clocks and Metabolism. Endocr Rev **35**: 648-670.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**: 1531–1545

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol **7**: 601–620

Frund J, Dormann CF, Tscharntke T (2011) Linne's floral clock is slow without pollinators--flower closure and plant-pollinator interaction webs. Ecol Lett **14**: 896–904

Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet **13**: 840–852

Futhcher B (2002) Transcriptional regulatory networks and the yeast cell cycle. Curr Opin Cell Biol **14**: 676–683

Geeven G, van Kesteren RE, Smit AB, de Gunst MCM (2012) Identification of context-specific gene regulatory networks with GEMULA--gene expression modeling using LAsso. Bioinformatics **28**: 214–221

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res **43**: D1049--56

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5**: R80

Glass L, Kauffman SA (1973) The logical analysis of continuous, non-linear biochemical control networks. J Theor Biol **39**: 103–129

Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, Aoki K, et al (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J **55**: 526–542

Goodspeed D, Chehab EW, Min-Venditti A, Braam J, Covington MF (2012) Arabidopsis synchronizes jasmonate-mediated defense with insect circadian behavior. Proc Natl Acad Sci U S A **109**: 4674–4677

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet **17**: 333–351

Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E (2013) On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol **5**: 578–590

Grewal SIS, Elgin SCR (2002) Heterochromatin: new possibilities for the inheritance of structure. Curr Opin Genet Dev **12**: 178–187

Grossman AR, Croft M, Gladyshev VN, Merchant SS, Posewitz MC, Prochnik S, Spalding MH (2007) Novel metabolism in Chlamydomonas through the lens of genomics. Curr Opin Plant Biol **10**: 190–198

Gualberti G, Papi M, Bellucci L, Ricci I, Bouchez D, Camilleri C, Costantino P, Vittorioso P (2002) Mutations in the Dof zinc finger genes DAG2 and DAG1 influence with opposite effects the germination of Arabidopsis seeds. Plant Cell **14**: 1253–1263

Guidi M, Ruault M, Marbouty M, Loiodice I, Cournac A, Billaudeau C, Hocher A, Mozziconacci J, Koszul R, Taddei A (2015) Spatial reorganization of telomeres in long-lived quiescent cells. Genome Biol **16**: 206

Haldar S, Saini A, Nanda JS, Saini S, Singh J (2011) Role of Swi6/HP1 self-association-mediated recruitment of Clr4/Suv39 in establishment and maintenance of heterochromatin in fission yeast. J Biol Chem **286**: 9308–9320

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor Newsl **11**: 10–18

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol **148**: 993–1003

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne J-B, Reynolds DB, Yoo J, et al (2004) Transcriptional regulatory code of a eukaryotic genome. Nature **431**: 99–104

Harmer SL, Hogenesch JB, Straume M, Chang HS, Han B, Zhu T, Wang X, Kreps JA, Kay SA (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. Science **290**: 2110–2113

Harmer SL, Kay SA (2005) Positive and negative factors confer phase-specific circadian regulation of transcription in Arabidopsis. Plant Cell **17**: 1926–1940

Harris EH (2001) Chlamydomonas as a model organism. Annu Rev Plant Physiol Plant Mol Biol **52**: 363–406

Henriksen PA, Kotelevtsev Y (2002) Application of gene expression profiling to cardiovascular disease. Cardiovasc Res **54**: 16–24

Ho Y, Costanzo M, Moore L, Kobayashi R, Andrews BJ (1999) Regulation of transcription at the Saccharomyces cerevisiae start transition by Stb1, a Swi6-binding protein. Mol Cell Biol **19**: 5267–5278

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods **9**: 473–476

Hoheisel JD (2006) Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet **7**: 200–210

Hong T, Watanabe K, Ta CH, Villarreal-Ponce A, Nie Q, Dai X (2015) An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. PLoS Comput Biol **11**: e1004569

Honkela A, Girardot C, Gustafson EH, Liu Y-H, Furlong EEM, Lawrence ND, Rattray M (2010) Model-based method for transcription factor target identification with limited data. Proc Natl Acad Sci U S A **107**: 7793–7798

Hu Q, Sommerfeld M, Jarvis E, Ghirardi M, Posewitz M, Seibert M, Darzins A (2008) Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. Plant J **54**: 621–639

Hughes JD, Estep PW, Tavazoie S, Church GM (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol **296**: 1205–1214

Hughes ME, Hogenesch JB, Kornacker K (2010) JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. J Biol Rhythm **25**: 372–380

Hwang S, Herrin DL (1994) Control of lhc gene transcription by the circadian clock in Chlamydomonas reinhardtii. Plant Mol Biol **26**: 557–569

Iliev D, Voytsekh O, Schmidt EM, Fiedler M, Nykytenko A, Mittag M (2006) A heteromeric RNA-binding protein is involved in maintaining acrophase and period of the circadian clock. Plant Physiol **142**: 797–806

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

Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature **409**: 533–538

Jacobshagen S, Johnson CH (1994) Circadian rhythms of gene expression in Chlamydomonas reinhardtii: circadian cycling of mRNA abundances of cab II, and possibly of beta-tubulin and cytochrome c. Eur J Cell Biol **64**: 142–152

Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, Myasnikova E, Surkova S, Vanario-Alonso CE, Samsonova M, Sharp DH, et al (2004) Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. Genetics **167**: 1721–1737

Jarolim S, Ayer A, Pillay B, Gee AC, Phrakaysone A, Perrone GG, Breitenbach M, Dawes IW (2013) Saccharomyces cerevisiae genes involved in survival of heat shock. G3 **3**: 2321–2333

Jaskowiak PA, Campello RJGB, Costa IG (2014) On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics **15 Suppl 2**: S2

Jiang W-K, Liu Y-L, Xia E-H, Gao L-Z (2013) Prevalent role of gene features in determining evolutionary fates of whole-genome duplication duplicated genes in flowering plants. Plant Physiol **161**: 1844–1861

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al (2011) Ancestral polyploidy in seed plants and angiosperms. Nature **473**: 97–100

Jimenez-Gomez JM, Corwin JA, Joseph B, Maloof JN, Kliebenstein DJ (2011) Genomic analysis of QTLs and genes altering natural variation in stochastic noise. PLoS Genet **7**: e1002295

Jiménez-Gómez JM, Wallace AD, Maloof JN (2010) Network analysis identifies ELF3 as a QTL for the shade avoidance response in Arabidopsis. PLoS Genet **6**: e1001100

Jin J, Zhang H, Kong L, Gao G, Luo J (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res **42**: D1182--7

Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature **527**: 384–388

Jones RF (1970) Physiological and Biochemical Aspects of Growth and Gametogenesis in Chlamydomonas-Reinhardtii. Ann N Y Acad Sci **175**: 648

Juven-Gershon T, Hsu J-Y, Theisen JW, Kadonaga JT (2008) The RNA polymerase II core promoter - the gateway to transcription. Curr Opin Cell Biol **20**: 253–259

Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol **9**: 770–780

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30**: 3059–3066

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **30**: 772–780

Kauffman S, Peterson C, Samuelsson B, Troein C (2003) Random Boolean network models and the yeast transcriptional network. Proc Natl Acad Sci U S A **100**: 14796–14799

Kazemian M, Pham H, Wolfe SA, Brodsky MH, Sinha S (2013) Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. Nucleic Acids Res **41**: 8237–8252

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428**: 617–624

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A **111**: 6131–6138

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al (2014) Reply to Brunet and Doolittle: Both selected effect and causal role elements can influence human biology and disease. Proc Natl Acad Sci U S A **111**: E3366

Keng T (1992) HAP1 and ROX1 form a regulatory pathway in the repression of HEM13 transcription in Saccharomyces cerevisiae. Mol Cell Biol **12**: 2616–2623

Kerr G, Ruskin HJ, Crane M, Doolan P (2008) Techniques for clustering gene expression data. Comput Biol Med **38**: 283–293

Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J **50**: 347–363

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol **14**: R36

Kinmonth-Schultz HA, Golembeski GS, Imaizumi T (2013) Circadian clock-regulated physiological outputs: dynamic responses in nature. Semin Cell Dev Biol **24**: 407–413

Kojima S, Shingle DL, Green CB (2011) Post-transcriptional control of circadian rhythms. J Cell Sci **124**: 311–320

Koranda M, Schleiffer A, Endler L, Ammerer G (2000) Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. Nature **406**: 94–98

Kuang Z, Boeke JD, Canzar S (2017) The dynamic landscape of fission yeast meiosis alternative-splice isoforms. Genome Res **27**: 145–156

Kucho K, Okamoto K, Tabata S, Fukuzawa H, Ishiura M (2005) Identification of novel clock-controlled genes by cDNA macroarray analysis in Chlamydomonas reinhardtii. Plant Mol Biol **57**: 889–906

Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. Nucleic Acids Res **34**: D74--81

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA (2012) Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci U S A **109**: 19498–19503

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res **22**: 1813–1831

Laporte D, Courtout F, Tollis S, Sagot I (2016) Quiescent Saccharomyces cerevisiae forms telomere hyperclusters at the nuclear membrane vicinity through a multifaceted mechanism involving Esc1, the Sir complex, and chromatin condensation. Mol Biol Cell **27**: 1875–1884

Lee TI, Young RA (2000) Transcription of eukaryotic protein-coding genes. Annu Rev Genet **34**: 77–137

Lee T-H, Tang H, Wang X, Paterson AH (2013) PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res **41**: D1152--8

Lehmann M, Gustav D, Galizia CG (2011) The early bee catches the flower - circadian rhythmicity influences learning performance in honey bees, Apis mellifera. Behav Ecol Sociobiol **65**: 205–215

Lehti-Shiu M, Panchy N, Wang P, Uygun S, Shiu S-H (2016) Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. BBA **1860**: 3–20

Lelli KM, Slattery M, Mann RS (2012) Disentangling the many layers of eukaryotic transcriptional regulation. Annu Rev Genet **46**: 43–68

Lespinet O, Wolf YI, Koonin E V, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res **12**: 1048–1059

Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. Proc Natl Acad Sci U S A **101**: 4781–4786

Li M, Hada A, Sen P, Olufemi L, Hall MA, Smith BY, Forth S, McKnight JN, Patel A, Bowman GD, et al (2015) Dynamic regulation of transcription factors by nucleosome remodeling. Elife 4

Li S, Brazhnik P, Sobral B, Tyson JJ (2008) A quantitative study of the division cycle of Caulobacter crescentus stalked cells. PLoS Comput Biol **4**: e9

Li Y, Chen C-Y, Kaye AM, Wasserman WW (2015) The identification of cis-regulatory elements: A review from a machine learning perspective. Biosystems **138**: 6–17

Li Y, Chen C-Y, Wasserman WW (2016) Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters. J Comput Biol **23**: 322–336

Li Y, Liang M, Zhang Z (2014) Regression analysis of combined gene expression regulation in acute myeloid leukemia. PLoS Comput Biol **10**: e1003908

Li Z, Defoort J, Tasdighian S, Maere S, de Peer Y, De Smet R (2016) Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. Plant Cell **28**: 326–344

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet **16**: 321–332

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Med Image Anal **42**: 60–88

Litt A, Irish V (2003) Duplication and Diversification in the APETALA1/FRUITFULL Floral Homeotic Gene Lineage: Implications for the Evolution of Floral Development. Genetics **165**: 821–833

Liu B, Hsu W, Ma Y (1998) Integrating Classification and Association Rule Mining. Proc. Fourth Int. Conf. Knowl. Discov. Data Min.

Liu H, Smith TPL, Nonneman DJ, Dekkers JCM, Tuggle CK (2017) A high-quality annotated transcriptome of swine peripheral blood. BMC Genomics **18**: 479

Liu M-J, Seddon AE, Tsai ZT-Y, Major IT, Floer M, Howe GA, Shiu S-H (2015) Determinants of nucleosome positioning and their influence on plant gene expression. Genome Res **25**: 1182–1195

Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H (2015) Characteristics of Plant Essential Genes Allow for within- and between-Species Prediction of Lethal Mutant Phenotypes. Plant Cell **27**: 2133–2147

Lomberk G, Bensi D, Fernandez-Zapico ME, Urrutia R (2006) Evidence for the existence of an HP1-mediated subcode within the histone code. Nat Cell Biol **8**: 407–415

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science (80-) **290**: 1151–1155

Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genomics **3**: 35–44

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, et al (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol **148**: 1772–1781

Macneil LT, Walhout AJM (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. Genome Res **21**: 645–657

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A **102**: 5454–5459

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, DREAM5 Consortium, Kellis M, Collins JJ, et al (2012) Wisdom of crowds for robust gene network inference. Nat Methods **9**: 796–804

Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G (2010) Revealing strengths and weaknesses of methods for gene network inference. Proc Natl Acad Sci U S A **107**: 6286–6291

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7 Suppl 1**: S7

Matsuo T, Ishiura M (2010) New insights into the circadian clock in Chlamydomonas. Int Rev Cell Mol Biol **280**: 281–314

McCarthy E, Mohamed A, Litt A (2015) Functional Divergence of APETALA1 and FRUITFULL is due to Changes in both Regulation and Coding Sequence. Front Plant Sci **6**: 1076

McNabb SL, Truman JW (2008) Light and peptidergic eclosion hormone neurons stimulate a rapid eclosion response that masks circadian emergence in Drosophila. J Exp Biol **211**: 2263–2274

Michael TP, McClung CR (2002) Phase-specific circadian clock regulatory elements in Arabidopsis. Plant Physiol **130**: 627–638

Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, Hazen SP, Shen R, Priest HD, Sullivan CM, et al (2008) Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genet **4**: e14

Miller JA, Widom J (2003) Collaborative competition mechanism for gene activation in vivo. Mol Cell Biol **23**: 1623–1632

Min S, Lee B, Yoon S (2016) Deep learning in bioinformatics. Brief. Bioinform.

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res **41**: e121

Mittag M, Kiaulehn S, Johnson CH (2005) The circadian clock in Chlamydomonas reinhardtii. What is it for? What is it similar to? Plant Physiol **137**: 399–409

Moghe GD, Hufnagel DE, Tang H, Xiao Y, Dworkin I, Town CD, Conner JK, Shiu S-H (2014) Consequences of Whole-Genome Triplication as Revealed by Comparative Genomic Analyses of the Wild Radish Raphanus raphanistrum and Three Other Brassicaceae Species. Plant Cell **26**: 1925–1937

Moghe GD, Shiu S-H (2014) The causes and molecular consequences of polyploidy in flowering plants. Ann N Y Acad Sci **1320**: 16–34

Monfared M, Simon M, Meister R, Roin-Villanova I, Kooiker M, Colombo L, Fletcher J, Gasser C (2011) Overlapping and antagonistic activities of BASIC PENTACYSTEINE genes affect a range of developmental processes in Arabidopsis. Plant J **66**: 1020–1031

Monnier A, Liverani S, Bouvet R, Jesson B, Smith JQ, Mosser J, Corellou F, Bouget FY (2010) Orchestrated transcription of biological processes in the marine picoeukaryote Ostreococcus exposed to light/dark cycles. BMC Genomics **11**: 192

Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N, Van Bel M, Poulain J, Katinka M, Hohmann-Marriott MF, et al (2012) Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. Genome Biol **13**: R74

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al (2014) The genome of Eucalyptus grandis. Nature **510**: 356–362

Nachman I, Regev A, Friedman N (2004) Inferring quantitative models of regulatory networks from expression data. Bioinformatics **20 Suppl 1**: i248--56

Nica AC, Dermitzakis ET (2013) Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci **368**: 20120362

Nolan T, Hands RE, Bustin SA (2006) Quantification of mRNA using real-time RT-PCR. Nat Protoc **1**: 1559–1582

Oakley T, Østman B, Wilson A (2006) Repression and loss of gene expression outpaces activation and gain in recently duplicated fly genes. Proc Natl Acad Sci U S A **103**: 11637–11641

Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, New York

Olson BJ, Oberholzer M, Li Y, Zones JM, Kohli HS, Bisova K, Fang SC, Meisenhelder J, Hunter T, Umen JG (2010) Regulation of the Chlamydomonas cell cycle by a stable, chromatin-associated retinoblastoma tumor suppressor complex. Plant Cell **22**: 3331–3347

O'Malley RC, Huang S-SC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165**: 1280–1292

Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, Achas M, Adebiyi E (2016) Clustering Algorithms: Their Application to Gene Expression Data. Bioinform Biol Insights **10**: 237–253

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol **53**: 673–684

Pajuelo E, Pajuelo P, Clemente MT, Marquez AJ (1995) Regulation of the expression of ferredoxin-nitrite reductase in synchronous cultures of Chlamydomonas reinhardtii. Biochim Biophys Acta **1249**: 72–78

Panchy N, Lehti-Shiu M, Shiu S-H (2016) Evolution of Gene Duplication in Plants. Plant Physiol **171**: 2294–2316

Panchy N, Wu G, Newton L, Tsai C-H, Chen J, Benning C, Farré EM, Shiu S-H (2014) Prevalence, evolution, and cis-regulation of diel transcription in Chlamydomonas reinhardtii. G3 **4**: 2461–2471

Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesch JB (2002) Coordinated transcription of key pathways in the mouse by the circadian clock. Cell **109**: 307–320

Panopoulou G, Hennig S, Groth D, Krause A, Poustka AJ, Herwig R, Vingron M, Lehrach H (2003) New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res **13**: 1056–1066

Pett JP, Korenčič A, Wesener F, Kramer A, Herzel H (2016) Feedback Loops of the Mammalian Circadian Clock Constitute Repressilator. PLoS Comput Biol **12**: e1005266

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res **21**: 447–455

Price C, Nasmyth K, Shuster T (1991) A general approach to the isolation of cell cycle-regulated genes in the budding yeast, Saccharomyces cerevisiae. J Mol Biol **218**: 543–556

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, et al (2012) The Pfam protein families database. Nucleic Acids Res **40**: D290-301

Qian W, Liao B, Chang A, Zhang J (2011) Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet **26**: 425–430

Ral JP, Colleoni C, Wattebled F, Dauvillee D, Nempont C, Deschamps P, Li ZY, Morell MK, Chibbar R, Purton S, et al (2006) Circadian clock regulation of starch metabolism establishes GBSSI as a major contributor to amylopectin synthesis in Chlamydomonas reinhardtii. Plant Physiol **142**: 305–317

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol **14**: R95

Ratcliff WC, Herron MD, Howell K, Pentz JT, Rosenzweig F, Travisano M (2013) Experimental evolution of an alternating uni- and multicellular life cycle in Chlamydomonas reinhardtii. Nat Commun **4**: 2742

Reimand J, Vaquerizas JM, Todd AE, Vilo J, Luscombe NM (2010) Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. Nucleic Acids Res **38**: 4768–4777

Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, Page JT, Udall JA, Wang X, Paterson AH, Wendel JF (2014) Ancient gene duplicates in Gossypium (cotton) exhibit near-complete expression divergence. Genome Biol Evol **6**: 559–571

Reuter JA, Spacek D V, Snyder MP (2015) High-throughput sequencing technologies. Mol Cell **58**: 586–597

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–140

Rodriguez J, Tang CHA, Khodor YL, Vodala S, Menet JS, Rosbash M (2013) Nascent-Seq analysis of Drosophila cycling gene expression. Proc Natl Acad Sci U S A **110**: E275–E284

Romanowski A, Yanovsky MJ (2015) Circadian rhythms and post-transcriptional regulation in higher plants. Front Plant Sci **6**: 437

Romero JM, Valverde F (2009) Evolutionarily conserved photoperiod mechanisms in plants: when did plant photoperiodic signaling appear? Plant Signal Behav **4**: 642–644

Rosa BA, Jiao YH, Oh S, Montgomery BL, Qin WS, Chen J (2012) Frequency-based time-series gene expression recomposition using PRIISM. Bmc Syst Biol. doi: Artn 69 Doi 10.1186/1752-0509-6-69

Sánchez L, Thieffry D (2001) A logical analysis of the Drosophila gap-gene system. J Theor Biol **211**: 115–141

Sanderson MJ, Thorne JL, Wikstrom N, Bremer K (2004) Molecular evidence on plant divergence times. Am J Bot **91**: 1656–1665

Santopolo S, Boccaccini A, Lorrai R, Ruta V, Capauto D, Minutello E, Serino G, Costantino P, Vittorioso P (2015) DOF AFFECTING GERMINATION 2 is a positive regulator of light-mediated seed germination and is repressed by DOF AFFECTING GERMINATION 1. Plant Biol **15**: 72

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet **37**: 501–506

Schnable J, Springer N, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A **108**: 4069–4074

Schnell S (2014) Validity of the Michaelis-Menten equation--steady-state or reactant stationary assumption: that is the question. FEBS J **281**: 464–472

Schranz M, Quijada P, Sung S, Lukens L, Amasino R, Osborn T (2002) Characterization and Effects of the Replicated Flowering Time Gene FLC in Brassica rapa. Gemetics **3**: 1457–1468

Schulze A, Downward J (2001) Navigating gene expression using microarrays--a technology review. Nat Cell Biol **3**: E190--5

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature **451**: 535–540

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet **20**: 461–464

Shi T, Ilikchyan I, Rabouille S, Zehr JP (2010) Genome-wide analysis of diel gene expression in the unicellular N(2)-fixing cyanobacterium Crocosphaera watsonii WH 8501. ISME J **4**: 621–632

Shimada A, Dohke K, Sadaie M, Shinmyozu K, Nakayama J-I, Urano T, Murakami Y (2009) Phosphorylation of Swi6/HP1 regulates transcriptional gene silencing at heterochromatin. Genes Dev **23**: 18–23

Shiu S-H, Shih M-C, Li W-H (2005) Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol **139**: 18–26

Shmulevich I, Dougherty ER, Kim S, Zhang W (2002) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics **18**: 261–274

Shmulevich I, Gluhovsky I, Hashimoto RF, Dougherty ER, Zhang W (2003) Steady-state analysis of genetic regulatory networks modelled by probabilistic boolean networks. Comp Funct Genomics **4**: 601–608

Siaut M, Cuine S, Cagnon C, Fessler B, Nguyen M, Carrier P, Beyly A, Beisson F, Triantaphylides C, Li-Beisson Y, et al (2011) Oil accumulation in the model green alga Chlamydomonas reinhardtii: characterization, variability between common laboratory strains and relationship with starch reserves. BMC Biotechnol **11**: 7

Sidorova JM, Mikesell GE, Breeden LL (1995) Cell cycle-regulated phosphorylation of Swi6 controls its nuclear localization. Mol Biol Cell **6**: 1641–1658

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al (2016) Mastering the game of Go with deep neural networks and tree search. Nature **529**: 484–489

Singh R, Lanchantin J, Robins G, Qi Y (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics **32**: i639--i648

Sinha S, Tompa M (2003) YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res **31**: 3586–3588

Soltis D, Bell C, Kim S, Soltis P (2008) Origin and early evolution of angiosperms. Ann N Y Acad Sci **1133**: 3–25

Soltis DE, Visger CJ, Soltis PS (2014) The polyploidy revolution then…and now: Stebbins revisited. Am J Bot **101**: 1057–1078

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics **14**: 91

Song YH, Ito S, Imaizumi T (2013) Flowering time regulation: photoperiod- and temperature-sensing in leaves. Trends Plant Sci **18**: 575–583

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell **9**: 3273–3297

Spitz F, Furlong EEM (2012) Transcription factors: from enhancer binding to developmental control. Nat Rev Genet **13**: 613–626

Spivak AT, Stormo GD (2016) Combinatorial Cis-regulation in Saccharomyces Species. G3 **6**: 653–667

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**: 2688–2690

Stamatakis A (2014) RAxMLversion 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**: 1312–1313

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the "Perceptron" algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res **10**: 2997–3011

Straume M (2004) DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. Methods Enzym **383**: 149–166

Sukumaran S, Almon RR, DuBois DC, Jusko WJ (2010) Circadian rhythms in gene expression: Relationship to physiology, disease, drug disposition and drug action. Adv Drug Deliv Rev **62**: 904–917

Takagi Y, Matsuda H, Taniguchi Y, Iwaisaki H (2014) Predicting the phenotypic values of physiological traits using SNP genotype and gene expression data in mice. PLoS One **9**: e115532

Takasaki H, Maruyama K, Takahashi F, Fujita M, Yoshida T, Nakashima K, Myouga F, Toyooka K, Yamaguchi-Shinozaki K, Shinozaki K (2015) SNAC-As, stress-responsive NAC transcription factors, mediate ABA-inducible leaf senescence. Plant J **84**: 1114–1123

Ter Linde JJM, Steensma HY (2002) A microarray-assisted screen for potential Hap1 and Rox1 target genes in Saccharomyces cerevisiae. Yeast **19**: 825–840

Teramoto H, Nakamori A, Minagawa J, Ono TA (2002) Light-intensity-dependent expression of Lhc gene family encoding light-harvesting chlorophyll-a/b proteins of photosystem II in Chlamydomonas reinhardtii. Plant Physiol **130**: 325–333

Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res **16**: 934–946

Thomas R (1973) Boolean formalization of genetic control circuits. J Theor Biol **42**: 563–585

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, et al (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol **3**: RESEARCH0088

Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, et al (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science **294**: 2364–2368

Trabzuni D, United Kingdom Brain Expression Consortium (UKBEC), Thomson PC (2014) Analysis of gene expression data using a linear mixed model/finite mixture model approach: application to regional differences in the human brain. Bioinformatics **30**: 1555–1561

Trairatphisan P, Wiesinger M, Bahlawane C, Haan S, Sauter T (2016) A Probabilistic Boolean Network Approach for the Analysis of Cancer-Specific Signalling: A Case Study of Deregulated PDGF Signalling in GIST. PLoS One **11**: e0156223

Tran L, Nakashima K, Sakuma Y, Simpson S, Fujita Y, Maruyama K, Fujita M, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2004) Isolation and functional analysis of Arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. Plant Cell **16**: 2481–2498

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**: 1105–1111

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol **28**: 511–515

Truenit E, Siemering K, Hdoge S, Grbic V, Haseloff J (2006) A Map of KNAT Gene Expression in the Arabidopsis Root. Plant Mol Biol **60**: 1–20

Truernit E, Haseloff J (2007) A Role for KNAT Class II Genes in Root Development. Plant Signal Behav **1**: 10–12

Tsai ZT-Y, Lloyd JP, Shiu S-H (2017) Defining Functional Genic Regions in the Human Genome through Integration of Biochemical, Evolutionary, and Genetic Evidence. Mol Biol Evol **34**: 1788–1798

Ueda HR, Hayashi S, Chen W, Sano M, Machida M, Shigeyoshi Y, Iino M, Hashimoto S (2005) System-level identification of transcriptional circuits underlying mammalian circadian clocks. Nat Genet **37**: 187–192

Uygun S, Peng C, Lehti-Shiu MD, Last RL, Shiu S-H (2016) Utility and Limitations of Using Gene Expression Data to Identify Functional Associations. PLoS Comput Biol **12**: e1005244

Uygun S, Seddon AE, Azodi CB, Shiu S-H (2017) Predictive Models of Spatial Transcriptional Response to High Salinity. Plant Physiol **174**: 450–464

van der Felden J, Weisser S, Brückner S, Lenz P, Mösch H-U (2014) The transcription factors Tec1 and Ste12 interact with coregulators Msa1 and Msa2 to activate adhesion and multicellular development. Mol Cell Biol **34**: 2283–2293

van Hoek MJA, Hogeweg P (2007) The role of mutational dynamics in genome shrinkage. Mol Biol Evol **24**: 2485–2494

van Noort V, Snel B, Huynen MA (2003) Predicting gene function by conserved co-expression. Trends Genet **19**: 238–242

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**: 530–536

Veitia RA, Bottani S, Birchler JA (2013) Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. Trends Genet **29**: 385–393

Voigt J, Munzner P (1987) The Chlamydomonas Cell-Cycle Is Regulated by a Light Dark-Responsive Cell-Cycle Switch. Planta **172**: 463–472

von Gromoff ED, Schroda M, Oster U, Beck CF (2006) Identification of a plastid response element that acts as an enhancer within the Chlamydomonas HSP70A promoter. Nucleic Acids Res **34**: 4767–4779

Voss TC, Hager GL (2014) Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Genet **15**: 69–81

Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S, et al (2012) The draft genome of a diploid cotton Gossypium raimondii. Nat Genet **44**: 1098–1103

Wang W, Haberer G, Gundlach H, Gläßer C, Nussbaumer T, Luo MC, Lomsadze A, Borodovsky M, Kerstetter RA, Shanklin J, et al (2014) The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun **5**: 3311

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

Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet **5**: 276–287

Weirauch M, Hughes T (2011) A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell Bichem **52**: 25–73

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

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, et al (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science **285**: 901–906

Wittenberg C, Reed SI (2005) Cell cycle-dependent transcription in yeast: promoters, transcription factors, and transcriptomes. Oncogene **24**: 2746–2755

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**: 708–713

Wong MKL, Krycer JR, Burchfield JG, James DE, Kuncic Z (2015) A generalised enzyme kinetic model for predicting the behaviour of complex biochemical systems. FEBS Open Bio **5**: 226–239

Woodhouse M, Tang H, Freeling M (2011) Different Gene Families in Arabidopsis thaliana Transposed in Different Epochs and at Different Frequencies throughout the Rosids. Plant Cell **23**: 4241–4253

Wu G, Hufnagel DE, Denton AK, Shiu S-H (2015) Retained duplicate genes in green alga Chlamydomonas reinhardtii tend to be stress responsive and experience frequent response gains. BMC Genomics **16**: 149

Xiang CC, Chen Y (2000) cDNA microarray technology and its applications. Biotechnol Adv **18**: 35–46

Yanovsky MJ, Kay SA (2002) Molecular basis of seasonal time measurement in Arabidopsis. Nature **419**: 308–312

Yeung MKS, Tegnér J, Collins JJ (2002) Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci U S A **99**: 6163–6168

Yin L, Huang C-H, Ni J (2006) Clustering of gene expression data: performance and similarity analysis. BMC Bioinformatics **7 Suppl 4**: S19

Yuan Y, Guo L, Shen L, Liu JS (2007) Predicting gene expression from sequence: a reexamination. PLoS Comput Biol **3**: e243

Yuh CH, Bolouri H, Davidson EH (2001) Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. Development **128**: 617–629

Zhang J, Tian X, Zhang H, Teng Y, Li R, Bai F, Elankumaran S, Xing J (2014) TGF-β-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. Sci Signal **7**: ra91

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol **9**: R137

Zhang Z, Belcram H, Magdelenat G, Couloux A, Samain S, Gill S, Rasmussen JB, Barbe V, Faris JD, Zhang Z, et al (2012) Correction for Zhang et al., Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. Proc Natl Acad Sci **109**: 1353–1353

Zheng X, Spivey N, Zeng W, Liu P, Fu Z, Klessig D, He S, Dong X (2012) Coronatine promotes Pseudomonas syringae virulence in plants by activating a signaling cascade that inhibits salicylic acid accumulation. Cell Host Microbe **11**: 587–596

Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah M V, Radhakrishnan M, et al (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res **19**: 556–566

Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN, Futcher B (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. Nature **406**: 90–94

Zones JM, Blaby IK, Merchant SS, Umen JG (2015) High-Resolution Profiling of a Synchronized Diurnal Transcriptome from Chlamydomonas reinhardtii Reveals Continuous Cell and Metabolic Differentiation. Plant Cell **27**: 2743–2769

Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH (2009) Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana. PLoS Genet **5**: e1000581

Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu SH (2011) Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proc Natl Acad Sci U S A **108**: 14992–14997

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu S-H (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol **151**: 3–15