Spring 5-15-2016

Development and Application of Comparative Gene Co-expression Network Methods in Brachypodium distachyon

Henry David Priest
Washington University in St. Louis

Recommended Citation

Priest, Henry David, "Development and Application of Comparative Gene Co-expression Network Methods in Brachypodium distachyon" (2016). *Arts & Sciences Electronic Theses and Dissertations*. 757.
https://openscholarship.wustl.edu/art_sci_etds/757

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Todd Mockler, Chair
Tom Brutnell
Barak Cohen
Justin Fay
Elizabeth Haswell
Lucia Strader

Development and Application of Comparative Gene Co-expression Network Methods in *Brachypodium distachyon*

by
Henry D. Priest

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2016
St. Louis, Missouri

Table of Contents

List of Figures v

List of Tables vi

Acknowledgments vii

ABSTRACT OF THE DISSERTATION ix

Introduction 1

1.1 *Cis*-regulatory elements in plant cell signaling 1

1.1.1 Bioinformatic approaches to plant *cis*-element prediction 2

1.1.2 Tools for plant cis-element prediction 3

1.1.3 Case studies elucidating functions of cis-elements in plant cell signaling 4

1.1.4 Distinguishing *bona fide cis*-elements from genomic 'noise' 8

1.1.5 Conclusions and future directions 9

Analysis of global gene expression in Brachypodium distachyon reveals extensive network plasticity in response to abiotic stress 12

2.1 Introduction 12

2.2 Results 14

2.2.1 Overall Differential Expression Analysis 14

2.2.2 Network Analysis of Stress Response in Brachypodium 17

2.2.3 Functional Annotation and Promoter Analysis 21

2.2.4 Unknown Module Members 27

2.2.5 Network Plasticity 27

2.2.6 Stress Responsive Modules in Brachypodium Transcriptional Circuitry 30

2.2.7 Conserved Abiotic Stress Responses 31

2.3 Discussion 36

2.3.1 Conserved Modular Responses 37

2.3.2 Network Plasticity 38

2.3.3 Web Resources 42

2.4 Methods 43

2.4.1 Experimental Growth Conditions and Tissue Sampling 43

2.4.2 RNA Preparation, Labeled cDNA Synthesis, and Microarray Hybridization 43

2.4.3 Mapping of Probes 44

2.4.4 Microarray Data Analysis 44
2.4.5 Heatmap and Principal Component Analysis 45
2.4.6 GO Analysis and Transcription Factor Annotation 45
2.4.7 Network Analysis 46
2.4.8 Promoter Analysis 47
2.4.9 Network Plasticity Analysis 47
2.4.10 Undefined Module Member Genes 48
2.4.11 Accession Number 49
dGCNA: edge-wise differential gene co-expression network analysis 50
3.1 Introduction 50
3.1.1 Gene Co-expression Network Analysis 50
3.1.2 Network Comparison and Elasticity 51
3.2 Algorithm & Methods 54
3.2.2 GO-term enrichment analysis 60
3.2.3 Promoter Analysis 60
3.2.4 Individual GCN Construction 61
3.3 Results 61
3.3.1 Differential Gene Co-expression Network Analysis 63
3.3.2 Comparison of Elasticity Networks to Standard GCNs 64
3.3.3 Biological Meaning of Elasticity Networks 71
3.3.4 Comparison to DiffCoEx 77
3.4 Discussion 79
3.4.1 Novel Whole-Network Biological Insights Revealed by Elasticity Analysis 80
3.4.2 Cluster Level Analysis 80
3.4.3 Meaning of node-node interactions in elasticity network 81
3.4.4 Comparison to Similar Methods 82
3.5 Conclusions 83
Analysis of disparate experiments via direct network comparison – a proof of concept 86
4.1 Introduction 86
4.1.1 Analysis of large-scale transcriptome datasets 87
4.2 Results 88
4.2.1 Parameter and Cutoff Calculation 88

4.2.2 Distribution of Network Sizes 91

4.2.3 Probe-set Stability Analysis 93

4.2.4 Prediction of Transcription Factor Binding Substrates 97

4.3 Discussion 98

4.3.1 Elasticity networks differ greatly in size 99

4.3.2 Methods of identification of significant elasticity 100

4.3.3 Interaction of Disparate Datasets to Enrich Biological Signal 101

4.3.4 Relationship between network properties and molecular activity of transcription factor loci 101

4.4 Conclusion 106

4.5 Methods 107

4.5.1 Plant Growth, Microarray Dataset Generation and Normalization 107

4.5.2 Probeset inclusion criteria 109

4.5.3 Network Comparison 109

4.5.4 Yeast 1-Hybrid Assays 110

4.5.5 Promoter Analysis 111

4.5.6 Gene Ontology Analysis 111

Future Directions 112

References 117

Appendix A 132

List of Figures

Figure 1. Discovery of transcriptional regulatory networks. 9
Figure 2. Possible experimental steps after an over-represented potential cis-element has been identified. 11
Figure 3. Differential expression of Brachypodium distachyon genes in response to stress 15
Figure 4. Weighted gene co-expression network of *Brachypodium* stress responsive genes. 18
Figure 5. Expression profiles of modules as a function of time in each stress condition.. ... 20
Figure 6 Scatterplot of transcription factor/target gene correlations. 28
Figure 7 Overlaps between gene-gene edge sets.. 65
Figure 8. Edges identified by dGCNA are distinct those identified using previous approaches.. 66
Figure 9. Mean-normalized, RMA-normalized gene expression for two modules of genes undergoing significant elasticity. 73
Figure 10. Elasticity Network Parameter and Cutoff Selections 89
Figure 11. Gene-set Stability Analysis 94
Figure 12. Expression patterns of abiotic-stress responsive, cell-cycle related genes experiencing positive elasticity under loss of thermocycles. 96
Figure 13. Interaction network of tested transcription factors and putative target promoters. 98

List of Tables

Table 1. Selected web-based resources for cis-element bioinformatics.. 4
Table 2. Module membership and functional and regulatory enrichment. 22
Table 3. Specific GO terms uniquely enriched in a selection of network modules. 24
Table 4. Specific short DNA sequences found to be statistically enriched in the promoters of module member genes. 26
Table 5. Putative network plasticity present between all pairwise conditional comparisons. 30
Table 6. Gene Ontology Terms Found to be Statistically Over-represented Only in the Elasticity Networks produced by dGCNA. 70
Table 7. Enrichment of the Evening Element in the Immediate Neighborhood of AtCCA1 and AtLHY. 75
Table 8. Description of Input Datasets. 88
Table 9. Statistically enriched Gene Ontology terms of stress-responsive genes which experience elasticity under loss of thermocycles. 94

Acknowledgments

I must acknowledge the tireless advocacy and mentorship of Todd Mockler. On more than a couple of occasions during my tenure as his student, he took actions in my favor above and beyond the call of mentorship.

I acknowledge the support of Oregon State University's Computational and Genome Biology Initiative Fellowship.

In addition, I acknowledge the numerous sources of funding which supported the work of Todd Mockler, and thereby, my work, during my studies:

NSF Plant Genome Grant DBI 0605240

USDA NRI Grant 2008-01077

DOE Plant Feedstock Genomics for Bioenergy Grant DE-FG02-08-ER64630.

Henry Priest

Washington University in St. Louis

Dec 2015

For my wife.

Dedicated to my Parents.

ABSTRACT OF THE DISSERTATION

Development and Application of Comparative Gene Co-expression Network Methods in

Brachypodium distachyon

by

Henry D. Priest

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2015

Professor Todd Mockler, Chair

Gene discovery and characterization is a long and labor-intensive process. Gene co-expression network analysis is a long-standing powerful approach that can strongly enrich signals within gene expression datasets to predict genes critical for many cellular functions. Leveraging this approach with a large number of transcriptome datasets does not yield a concomitant increase in network granularity. Independently generated datasets that describe gene expression in various tissues, developmental stages, times of day, and environments can carry conflicting co-expression signals. The gene expression responses of the model C3 grass *Brachypodium distachyon* to abiotic stress is characterized by a co-expression-based analysis, identifying 22 modules of genes, annotated with putative DNA regulatory elements and functional terms. A great deal of co-expression elasticity is found among the genes characterized therein. An algorithm, dGCNA, designed to determine statistically significant changes in gene-gene co-expression relationships is presented. The algorithm is demonstrated on the very well-characterized circadian system of *Arabidopsis thaliana*, and identifies potential strong signals of molecular interactions between a specific transcription factor and putative target gene loci. Lastly, this network comparison approach based on edge-wise similarities is demonstrated on many pairwise comparisons of independent microarray datasets, to demonstrate the utility of fine-grained network comparison, rather than amassing as large a dataset as possible. This approach identifies a set of 182 gene loci which are differentially expressed under drought stress, change their co-expression strongly under loss of thermocycles or high-salinity stress, and are associated with cell-cycle and DNA replication functions. This set of genes provides excellent candidates for the generation of rhythmic growth under thermocycles in *Brachypodium distachyon.*

Introduction

1.1 *Cis*-regulatory elements in plant cell signaling

Cell signaling is one aspect of the complex system of communication that coordinates basic cellular activities and interactions of a cell with its environment. Transcriptional regulatory networks that drive organ and cell specific patterns of gene expression and mediate interactions with the environment represent one aspect of plant cell signaling. Fundamentally, the transcriptional regulation of gene expression in eukaryotes is mediated by recruitment of transcription factors (TFs) to *cis*-regulatory elements. Transcription factors interact with specific DNA elements, other transcription factors, and the basal transcriptional machinery to regulate the expression of target genes. In plants, transcriptional regulation is mediated by more than 1,500 TFs and each TF controls the expression of tens or even thousands of target genes in complex signaling networks [1,2]. Transcription factor binding sites (or '*cis*-elements'; 'motifs') are the functional DNA elements that influence temporal and spatial transcriptional activity. Multiple *cis*-elements comprise *cis*-regulatory modules (CRMs), which integrate signals from multiple TFs resulting in combinatorial control, and highly specific patterns of gene expression. Therefore, identifying and understanding the functions of *cis-element*s, and their combinatorial role in CRMs, is essential for elucidating the mechanisms by which cells perceive and correctly respond to their environment, and participate in organism development and homeostasis. With the recent availability of several high-quality sequenced and annotated plant genomes, large public databases of global gene expression measurements, and easy access to expression profiling technologies for individual laboratories, there has been a surge of studies involving transcription factor binding sites and their role as components of a larger transcriptional network. This review discusses relevant recent studies of plant *cis*-elements, focusing primarily on studies including prediction of *cis*-elements from high throughput expression profiling datasets and bioinformatics analysis of upstream sequences regulating co-expressed genes.

1.1.1 Bioinformatic approaches to plant *cis*-element prediction

Genome-wide expression profiling experiments have greatly facilitated *cis*-element prediction. To date, microarrays have been the most widely used platform used to measure steady state mRNA levels in plants. Groups of co-expressed genes are identified using microarrays, and assumed to be co-regulated. The presumed upstream regulatory regions of arbitrary length are then used to identify candidate DNA motifs. Multiple motifs that are over-represented in the promoters of a co-expressed gene cluster may represent the same CRM, and therefore be acting in a combinatorial mode. There are several obvious limitations with this approach. The underlying assumption – that co-expressed genes are transcriptionally co-regulated – may not always be true. Microarray assays measure steady state transcript levels in a particular sample, not transcriptional activity *per se*. Most microarray-based transcript profiling experiments cannot distinguish between changes in transcript levels caused by post-transcriptional regulation (i.e. transcript stability) rather than *cis*-element mediated transcriptional regulation. Most array assays are prone to ignore potential complications due to samples comprised of multiple tissue/cell types or changes in RNA content per cell. Moreover, the upstream regulatory sequences that are analyzed are arbitrary, and limited by the quality of the underlying genome sequence and its annotation. If a gene model is incorrectly annotated, the potential upstream regulatory sequence will be incorrect as well. Nevertheless, despite these potential limitations, in non-plant systems such as yeast many studies involving integration of data from transcript profiling, chromatin immunoprecipitation (ChIP) experiments, and genomic and computational transcription factor binding site predictions, have borne out the strength of the co-expression/co-regulation assumption [3–9]. Moreover, as described below (see Case Studies), several studies in plants have illustrated the utility of co-expression-driven prediction of *cis*-elements as a means to begin deciphering transcriptional networks.

1.1.2 Tools for plant cis-element prediction

A number of algorithms and bioinformatics tools have been developed to identify potential cis-elements in the regulatory sequences of co-expressed genes (reviewed in [10–12]). The fundamental assumptions underlying the computational approaches are that co-regulated genes should contain similar cis-elements in their upstream regulatory regions at statistically significant levels. Regardless of the exact algorithmic details, generically speaking, the computational approaches for identifying putative cis-elements estimate the probability of occurrences of short DNA motifs by comparing the observed number of occurrences of a particular motif in a set of sequences to the expected number of occurrences based on random sampling or statistical modeling of a background distribution [13–23]. Therefore each algorithm requires a background model to calculate the expected frequency for each motif. The composition of the sequences underlying the background model is critical because the various sequence features within a genome exhibit different base compositions. Moreover, as with gene prediction programs, background models must be generated on a species-by-species basis. At a minimum, this requires an available genome sequence and annotation. An overview of the available web-based tools for the identification and prediction of cis-elements in plants is provided in Table 1. It should be noted that there is a plethora of web-based tools for cis-element prediction that are not specific to plants, but generally applicable to plant studies - for example, MEME (http://meme.sdsc.edu) [24].

Table 1. Selected web-based resources for cis-element bioinformatics

Resource	**Type***	**URL**	**References**
AGRIS	D	http://arabidopsis.med.ohio-state.edu/	[169]
AtCOECis	D, P	http://bioinformatics.psb.ugent.be/ATCOECIS/	[36]
Athamap	D	http://www.athamap.de/	[170]
Athena	D, P	http://www.bioinformatics2.wsu.edu/cgi-bin/Athena/cgi/home.pl	[171]
BAR Promomer	P	http://bar.utoronto.ca/ntools/cgi-bin/BAR_Promomer.cgi	[23]
DATF	D	http://datf.cbi.pku.edu.cn/	[172]
DOOP	D	http://doop.abc.hu/	[173]
ELEMENT	P	http://element.cgrb.oregonstate.edu	[20,21]
Improbizer	P	http://users.soe.ucsc.edu/~kent/improbizer/improbizer.html	-
MEME	P	http://meme.sdsc.edu/meme4_1/intro.html	[174–176]
MotifSampler	P	http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html	[14,16]
PLACE	D	http://www.dna.affrc.go.jp/PLACE/	[177]
PlantCARE	D	http://bioinformatics.psb.ugent.be/webtools/plantcare/html/	[91]
PlantPAN	P, D	http://plantpan.mbc.nctu.edu.tw/	[178]
PlantProm DB	D	http://linux1.softberry.com/berry.phtml?topic=plantprom&group=data&subgroup=plantprom	[179]
Plant TF DB	D	http://planttfdb.cbi.pku.edu.cn/	[2]
Plant Promoter DB	D	http://ppdb.gene.nagoya-u.ac.jp/cgi-bin/index.cgi	[180]
RSAT	P	http://rsat.ccb.sickkids.ca/	[181]
TAIR pattern match	P	http://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl	[182]
Transfac	D	http://www.gene-regulation.com/pub/databases.html	[183]
WeederWeb	P	http://159.149.109.9/modtools/	[15,184]

* Type: D, Database; P, Prediction.

1.1.3 Case studies elucidating functions of cis-elements in plant cell signaling

In recent years, the rapid accumulation of genome-scale datasets – including genome sequences, genome annotations, gene-function predictions, and expression profiling experiments – has facilitated systems approaches aimed at discovery of cis-elements, and interrogating their roles in plant cell signaling. The proliferation of microarray experiments, in particular those designed for model plants with a high quality genome annotation such as Arabidopsis, provide whole genome catalogs of transcript levels. These microarray datasets represent different stages of development, organs, cell types, environmental conditions, and various other stimuli or treatments. Moreover, there has been a tremendous increase in the availability of so-called 'expression atlas' datasets in the public domain [21,25–27]. Here we briefly review several studies in recent years that have integrated expression and sequence data in order to identify cis-elements and information about their functions in plant cell signaling. Typically, differentially expressed genes were identified under some stimulus, treatment, or environmental condition, or by comparing different genotypes. The upstream regulatory regions of the differentially expressed genes were subjected to bioinformatic analyses to identify overrepresented cis-elements. Several studies have been focused on phytohormone signaling or stress response signaling. For example, analysis of upstream regulatory regions of genes coordinately regulated by treatments with auxin and brassinosteroid phytohormones revealed shared overrepresented cis-elements and novel crosstalk between phytohormone signaling pathways [19]. Analysis of promoters of Arabidopsis genes that were differentially expressed in plants treated with abscisic acid (ABA) or abiotic stresses (drought, cold, salt), identified a number of ACGT-containing ABA response elements (ABREs) and coupling elements [28]. In another study, microarrays were used to compare global gene expression responses of wild-type Arabidopsis plants and mutants defective in 'retrograde' signaling between the chloroplast and nuclear genomes [29]. Promoter analysis for genes de-repressed in retrograde signaling mutants revealed an over-represented ACGT motif, the core of both the light-responsive G-box (CACGTG) and ABREs, thus demonstrating a new connection between phytohormone, sugar, and retrograde signaling. Walley and colleagues [30] sought to elucidate the molecular mechanisms by which stress signals are transduced in plants. Mechanical wounding of Arabidopsis leaves was used as a stress stimulus and microarray analysis identified hundreds of wound responsive genes. Bioinformatic analysis identified a novel over-represented motif, CGCGTT, termed the Rapid Stress Response Element (RSRE), occurring in the promoters of genes upregulated during wounding stress. Subsequent experiments using luciferase reporter constructs and mutations in the RSRE motif demonstrated that the RSRE cis-element is sufficient to confer stress responsiveness in vivo. Using a microarray-driven approach Evrard et al.[31] identified FORCA - a hexameric cis-element that is conserved in clusters of Arabidopsis genes co-expressed in response to fungal pathogens and light treatments. It was proposed that the FORCA element integrates light- and defense-related signals in Arabidopsis and participates in the transcriptional adjustments to environmental changes.

A few recent studies [21,32,33] have identified novel components of diurnal/circadian transcriptional networks. Most eukaryotes use daily light/dark cycles as timing cues to ensure that a wide variety of biological processes are phased to occur at the correct times of day. In one study, a bioinformatics pipeline for discovery of transcriptional networks was applied to microarray datasets interrogating the transcriptomes of Arabidopsis plants grown in different light, temperature, or circadian conditions. Mining the promoters of cycling genes identified three cis-acting modules controlling time of day expression: the morning elements, comprising the morning CRM (ME, CCACAC)/G-box (CACGTG); the evening elements, comprising the evening CRM (EE, AAATATCT)/GATA (GATA); and the midnight elements, comprising the midnight CRM (TBX, AAACCCT)/starch synthesis box (SBX, AAGCCC)/ protein box (PBX, ATGCCC). These three modules are conserved across distantly related species such as Arabidopsis, rice, poplar, and papaya [21,34] suggesting that diurnal and circadian signaling have shaped the evolution of plant transcriptional networks and allow plants to adapt to diverse and ever-changing daily environments. In another study [33] aimed at elucidating the interactions between light signaling, the circadian clock, and growth-promoting phytohormone pathways in plant growth, a novel cis-element (HUD; CACATG) was over-represented in the promoters of plant hormone-associated genes that are co-expressed near dawn, the time of day when hypocotyl growth rate is maximal. The HUD element was shown to be sufficient to confer predicted diurnal and circadian expression patterns when used to drive expression of a luciferase reporter construct in vivo.

To date, several attempts have been made to extend the general approaches described above to the large public Arabidopsis expression atlas datasets. For example, Walther et al. [35] used the large AtGenExpress database (http://www.arabidopsis.org/info/expression/ATGenExpress.jsp) to test their hypothesis that genes differentially expressed in response to several different stimuli should contain a greater number of distinct cis-elements in their upstream regions than genes that respond to relatively few stimuli. By combining differential gene expression patterns with an analysis of cis-elements in Arabidopsis promoters, they found a positive correlation between genes that respond to multiple stimuli and the density of cis-elements in their upstream regions. Perhaps not surprisingly, genes predicted to function in the regulation of transcription, stress responses, and signaling processes exhibited the greatest regulatory capacity. In another study Vandepoele and co-workers [36] integrated predictions of CRMs, previously known and potential novel cis-elements, and predictions of gene function (e.g. GO annotations) to annotate ~9,100 clusters of co-expressed genes with potential cis-elements, including hundreds of evolutionarily conserved, but previously unknown, cis-elements. These annotations of over-represented cis-elements in co-expressed gene clusters provide powerful new resources for elucidating the mechanisms underlying transcriptional control in plants and inferring functional information for Arabidopsis genes.

1.1.4 Distinguishing *bona fide cis*-elements from genomic 'noise'

It remains challenging to distinguish potential cis-elements that serve as genuine transcription factor binding sites from genomic background noise. The canonical short palindromic 'G-box' (CACGTG) represents an illustrative example of such a challenge. The G-box is one of the best-studied cis-elements, and has been shown to drive gene expression in plants in response to light [37]. Several studies have shown that the G-box is frequently over-represented in the promoter sequences of certain co-expressed genes or in intragenomic conserved noncoding sequences (CNSs). For example, Freeling et al. [38,39] analyzed 14,944 Arabidopsis CNSs and demonstrated that many known TF binding motifs, including the G-box, are overrepresented in these CNSs. In our own studies of circadian and diurnal regulation of gene expression in Arabidopsis [20,21] we found the G-box to be overrepresented in the promoters of several hundred genes whose diurnal expression peaked a few hours after dawn in short-day photoperiod conditions. However, like other relatively short DNA motifs, the G-box occurs in all regions of plant genomes (in 'promoters', intergenic regions, coding regions, introns etc.). The G-box occurs in approximately 29,000 locations in the Arabidopsis genome, and occurs more often in annotated genic regions than in intergenic regions. Obviously, it would be naïve to expect every occurrence of the G-box to function as a transcription factor-binding site in vivo, regardless of its sequence context. This problem is even more exaggerated in the case of shorter motifs such as the ubiquitous GATA element, which occurs on average several times in every potential upstream regulatory region. Recent approaches based on the new high throughput sequencing technologies will greatly facilitate efforts to identify the functional instances of predicted cis-elements. Chromatin immunoprecipitation coupled with deep sequencing (ChIP-seq; [40,41]) can be used to identify individual transcription factor targets and whole-genome mappings of nucleosome locations can associate chromatin organization with transcriptional activity [42].

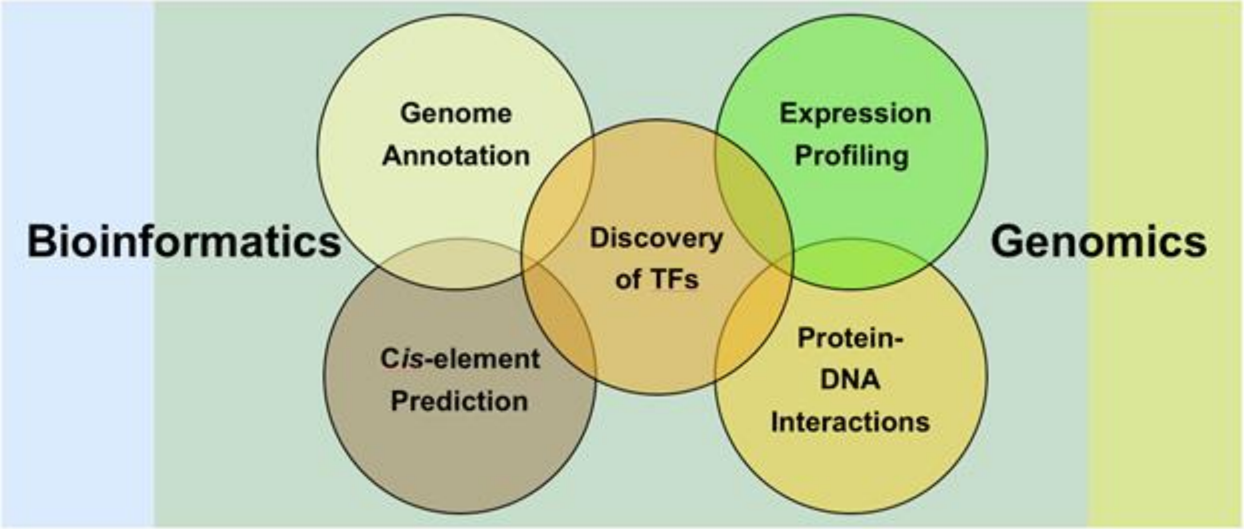


Figure 1. Discovery of transcriptional regulatory networks.
The elucidation of transcriptional regulatory networks is a holistic process involving both computational and experimental biology approaches that are interdependent and increasingly driven by high-throughput technologies. For example, *cis*-element discovery will be increasingly dependent on high-quality empirical genome annotations generated using advanced transcription unit assembly algorithms. Whole-genome expression profiling experiments and clustering of co-expressed genes, again driven by technology improvements will exhibit greater spatiotemporal resolution and sensitivity. High-throughput one-hybrid screens will facilitate identification of putative transcription factors that interact with *cis*-elements and promoters of interest. Whole-genome analyses of protein-DNA interactions, facilitated by HTS technologies, will identify *in vivo* transcription factor binding sites, chromatin modifications, and nucleosome positions, elucidating global regulatory networks.

1.1.5 Conclusions and future directions

The transcriptional control of gene expression depends on a balance between activating and repressing regulatory components in upstream regulatory regions. Cis-elements play a central role in gene regulation by integrating signals at the DNA level upstream of a target gene. Despite the fact that several recent studies have used high throughput genome scale datasets and bioinformatics approaches to elucidate cis-elements implicated in plant signaling, these are still the early days. We can reasonably expect that technological advances, such as 'digital gene expression' profiling (DGE; [43]) will make it possible to profile and map spatiotemporal gene expression more precisely, enabling finer clustering of co-expressed genes and better predictions of cis-elements. Recent advances in high-throughput transcriptome sequencing will facilitate better genome annotations and precise empirical annotations of transcriptional start sites, which will in turn yield better predictions of regulatory regions. New approaches based on high throughput sequencing enable acquisition of high-resolution global protein-DNA interaction maps. These maps can identify genuine functional transcription factor binding sites in vivo. The integration of global mappings of transcription factor binding sites and dynamic remodeling of nucleosomes with global expression profiling and cis-element predictions will provide the foundation for systematic reconstructions of gene regulatory networks (**Figure 1**). Moreover, new functional genomics approaches have been developed for identification of transcription factors that interact with cis-elements of interest - such as the recently developed high-throughput yeast-one-hybrid (Y1H) library screening system ("Promoter Hiking") [44]. Classic molecular and genetic approaches can be coupled with these newer high-throughput methods and bioinformatics in a series of rational experimental steps after identification of a predicted cis-element (Figure 2). Given that the vast majority of plant transcription factors remain unstudied and the cis-elements corresponding to most transcription factors are unknown, we can be certain there is still plenty of room for pioneers.

1.1.6 Publication Record and Author Contributions

This work was previously published as [45]. Henry Priest wrote the manuscript with revisions and contributions from Sergei Filichkin and Todd Mockler. Written permission for use of this material has been obtained from Sergei Filichkin and Todd Mockler.

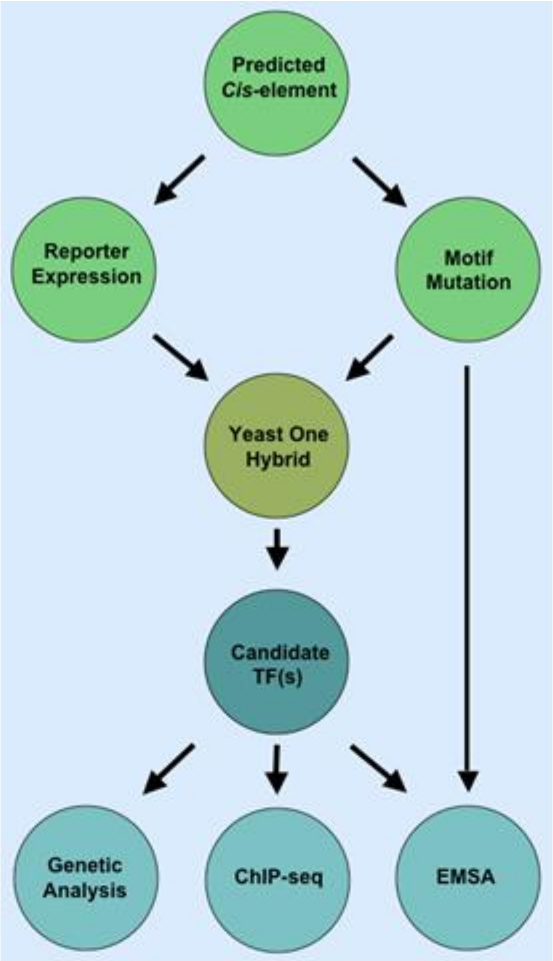


Figure 2. Possible experimental steps after an over-represented potential cis-element has been identified. Beginning with a predicted cis-element, the flowchart depicts a series of experimental steps that can be pursued to elucidate the function of the cis-element in transcriptional regulation. For example, recapitulation studies using intact and mutated versions of the predicted cis-element driving a reporter gene such as Luciferase can be used to validate its hypothesized function in vivo. One-hybrid screens (traditional or high-throughput) can be used to identify putative transcription factors that interact with the element of interest. After a transcription factor candidate is identified, molecular genetic analysis of mutants, and in planta over-expression and/or knockdown approaches can be used to functionally characterize the interacting transcription factor. Electrophoretic mobility shift assays (EMSA) can be used to confirm protein:DNA interactions in vitro. For example, expression-profiling approaches can be used to characterize molecular phenotypes in a transcription factor mutant, including mis-regulation of target genes. Finally, global analysis of protein-DNA interactions, for example using ChIP-seq, can be used to identify the in vivo transcription factor binding sites.

Analysis of global gene expression in *Brachypodium distachyon* reveals extensive network plasticity in response to abiotic stress

2.1 Introduction

Plants are sessile organisms that have evolved an exceptional ability to perceive, respond, and adapt to their environment. Environmental stresses are a major limiting factor in agricultural productivity [46,47], as plant growth is severely affected by environmental conditions such as cold, high-salinity, drought, and heat [48,49]. In comparison to *Arabidopsis thaliana* and *Oryza sativa*, relatively little is known about how many agriculturally important cereals (e.g., wheat, corn, barley) respond to abiotic stresses [50–53]. The stress-induced transcriptomic responses of plants reveal the molecular mechanisms underlying the abiotic stress response. An understanding of these mechanisms will allow researchers to improve stress tolerance of food crops to enhance agricultural productivity under imperfect growing conditions to ensure the world's long-term food security [54–56].

The abiotic stress response occurs in two stages, an initial sensory/activation stage followed by a physiological stage during which the plant responds to the perceived stress [48,57,58]. Once a stress cue is perceived, secondary messengers such as calcium and inositol phosphates [59] and reactive oxygen species (ROS) are produced. The increase in Ca^{2+} is sensed by various calcium-binding proteins that initiate phosphorylation cascades that subsequently activate transcription factors [60,61]. Transcription factors in turn activate expression of stress responsive genes. This begins the second phase and elicits physiological changes necessary to survive the particular environmental stress (reviewed in [58]). The genes expressed and subsequent physiological changes induced during the second phase are dependent upon the particular abiotic stress encountered. These changes can include modifications to cell membrane components – resulting in changes in membrane fluidity [62], stomatal closure [63], decreased photosynthetic activity [64,65], and increased production of heat shock proteins (HSPs) or dehydrin cryoprotectants [48].

Previous work in monocot stress responses has been completed in rice (*Oryza sativa* ssp. *japonica* cv. 'Nipponbare' and ssp. *indica* cv. 'Minghui 63'). Expression levels of 20,500 transcriptional units in rice callus treated with abscisic acid (ABA) and gibberellin were evaluated using oligonucleotide arrays [66]. A more comprehensive approach using a microarray querying 36,926 genes was used to profile expression responses of rice to drought and high-salinity stresses in three tissues [51]. Recently, profiling of transcriptional responses to cold stresses in winter barley was performed using a microarray-based approach [67], and the transcriptional responses of three wheat cultivars to cold stress was explored in a separate study using microarray-based approaches [68].

Here, we present a genome-wide survey of *Brachypodium* transcript-level gene expression responses to four abiotic stresses: heat, high salinity, drought, and cold. We found significant differences in responses of the *Brachypodium* transcriptome to the four abiotic stresses in terms of timing and magnitude. We were able to identify 22 modules, 10 of which defined clear biological processes. As expected from studies of other plant model systems, photosynthesis, cell cycle and cell wall expression modules were down-regulated under abiotic stress. We found that the modules up-regulated by salt and drought fell into unique gene ontology (GO) categories, whereas cold and heat up-regulated transcription factor (TF) expression and expression of genes involved in stabilizing protein folding, respectively. The response of *Brachypodium* to heat, high salinity, drought, and cold stress was profiled over twenty-four hours after the onset of stress conditions.

This study represents a significant development in genomics resources for *Brachypodium*, a close relative of many agriculturally and economically important cereal crop species.

2.1.1 Publication Record and Author Contributions

This work was previously published as [69]. The experiment was conceived and designed by Todd Mockler, Todd Michael, and Sam Fox. Tissue collection and RNA preparation was completed by Sam Fox and Jessica Murray. Henry Priest conceived and executed all analyses. Henry Priest, Sam Fox, Erik Rowley, and Todd Mockler wrote the manuscript. Written permission for use of this material has been obtained from all authors.

2.2 Results

2.2.1 Overall Differential Expression Analysis

Drought, high-salinity, cold, and heat are four important abiotic stresses that adversely affect the productivity of plants. We surveyed *Brachypodium* transcript-level gene expression responses to these stresses using the Affymetrix *Brachypodium* Genome Array (BradiAR1b520742). This microarray queries all annotated genes in the *Brachypodium* genome with multiple individual probes targeting each gene. The response of *Brachypodium* to heat, high salinity, drought, and cold stress was profiled in an asymmetric time-course over the twenty-four hours immediately following onset of stress conditions. This allowed us to monitor the transcriptional responses of the plant to stress rather than endogenous circadian or diurnal rhythms. Biological triplicate samples were taken from control and stressed plants at each time point.

Overall, 3,105 genes were significantly up-regulated and 6,763 genes were significantly down-regulated in response to at least one abiotic stress. In response to cold, heat, salt, and drought stresses 40, 1,621, 1,137, and 5,790 genes were significantly down-regulated, respectively. In contrast, 447, 458, 1,565, and 2,290 genes were significantly up-regulated in response to cold, heat, salt, and drought stress, respectively.

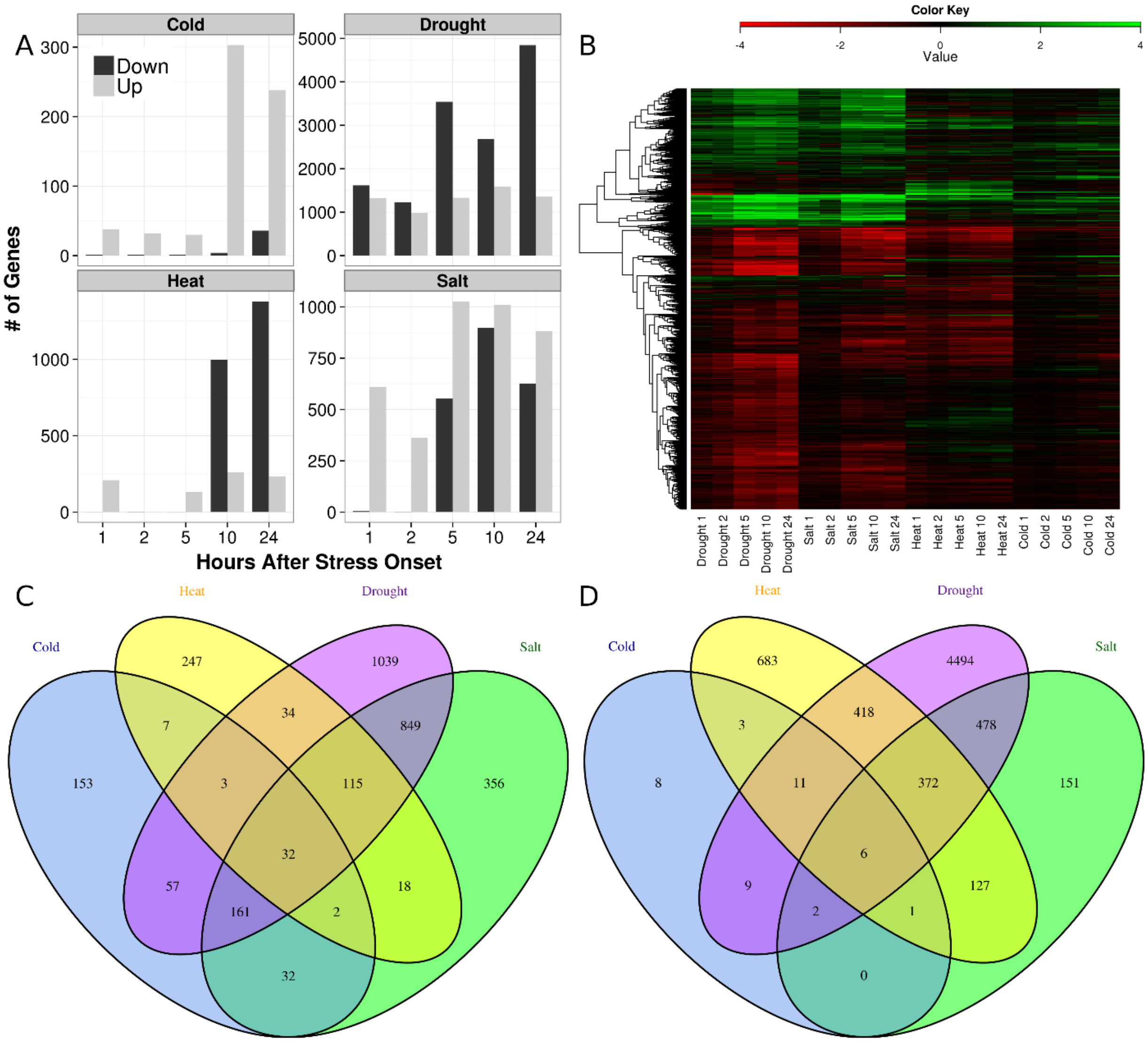


Figure 3. Differential expression of Brachypodium distachyon genes in response to stress. **A**. Numbers of genes up-regulated (light grey bars) and down-regulated (dark grey) are shown as a function of time in hours after stress onset. **B**. Heatmap of expression differences between control and indicated stress arrays. Similar expression profiles are clustered in the dendrogram. Positive (green) and negative (red) differences between stress and control arrays are shown for all genes called as differentially expressed by SAM analysis. Columns are time points. Expression values are saturated at +/- 4 RMA, for display purposes. **C.** Venn diagram showing overlap of up-regulated genes in response to the four assayed abiotic stresses: cold (blue), heat (yellow), drought (purple) and salt (green). Area of overlaps is not proportional to the overlap. The numbers of genes in each region of the diagram are indicated. **D.** Venn diagram depicting intersections of sets of down-regulated genes in response to the four assayed abiotic stresses

The overall number of genes differentially expressed in each stress condition increased over time (**Figure 3A**); the directionality of differential expression differed strikingly with the type of stress. The cold stress response consisted almost entirely of up-regulated genes; very few genes were down-regulated at twenty-four hours (**Figure 3A,** top left). In contrast, the response to heat stress was primarily down regulation (**Figure 3A,** bottom left). Up-regulation of certain genes in response to heat stress response was observed after 1 hour, but no significant differential expression was observed at 2 hours after onset of stress. After 10 and 24 hours of heat treatment, more than 1,000 genes were down-regulated. Between 1,000 and 2,000 genes were up-regulated at all time points of drought treatment (**Figure 3A**, top right). Down-regulation of genes was low in the early phases of drought response and increased drastically as the treatment continued beyond 2 hours. More than 2,500 genes were differentially expressed 5, 10, and 24 hours after drought onset. Early in the response to salt stress, only up-regulation of genes was observed. At 5 hours post-onset, down-regulation was observed in conjunction with up-regulation with neither as dominant as was seen in the other three stresses (**Figure 3A,** bottom right).

Drought and salt stresses yielded the most similar patterns of variance, whereas the cold and heat stress responses differed strongly from each of the other two stresses and from each other. Similarities were observed in the heatmap depicting hierarchical clustering of the expression data (**Figure 3B**) in which the Robust Multi-array Average (RMA) [70] expression value differences between mRNA abundances in control and stress-treated plants are plotted for all stress conditions. The overall similarity between the salt and drought stress responses can also be seen in this heatmap and is also reflected in the principal component analysis (PCA) results (**Supplemental Figure 1**).

A large number of genes are differentially expressed only under drought stress (purple ovals, **Figure 3C and Figure 3D**). In response to drought treatment, 1,039 genes were up-regulated and 4,494 were down-regulated. Only about half of the genes differentially expressed in the heat treatment were also responsive to drought (1,088 of 2,079 genes responsive to heat were also responsive to drought). Further, 44.7% of all genes differentially expressed in response to heat stress were unique to that response (930 of 2,079, compare yellow to purple ovals in **Figure 3C** and **Figure 3D**). Only about 25% of genes differentially expressed upon salt treatment were independent of the drought response (687 of 2,702), and even fewer were unique to salt (507 of 2,702, 18.8%; compare green to purple ovals in **Figure 3C** and **Figure 3D**). The response to extended cold treatment had strong overlap with the drought response as well. Only 206 genes were responsive to cold stress and not to drought treatment (206 of 487, 42.3%), and 161 genes (of 487 differentially regulated by cold relative to unstressed plants) were uniquely regulated by cold stress (compare blue to purple ovals, **Figure 3C** and **Figure 3D**). From these analyses, the complex nature of the timing of gene regulation in response to stresses (**Figure 3**), the differences in intensities of differential expression in response to stresses (**Figure 3B**), and the extensive overlap among genes regulated during stress responses (**Figure 3C and Figure 3D**) are apparent.

2.2.2 Network Analysis of Stress Response in Brachypodium

In order to further analyze the systematic transcriptional responses of Brachypodium to abiotic stresses, we performed weighted gene co-expression network analysis (WGCNA) on data collected on the 9,496 differentially expressed genes using the WGCNA package in R [25]. Gene modules are composed of genes that share similar profiles and have high correlations with each other. The weighted interaction network is shown in **Figure 4**. Nodes (genes) are connected by edges (co-expression relationships). The connection between two nodes was determined by the correlation between the expression levels of the genes those nodes represent across all experiments used in the analysis.

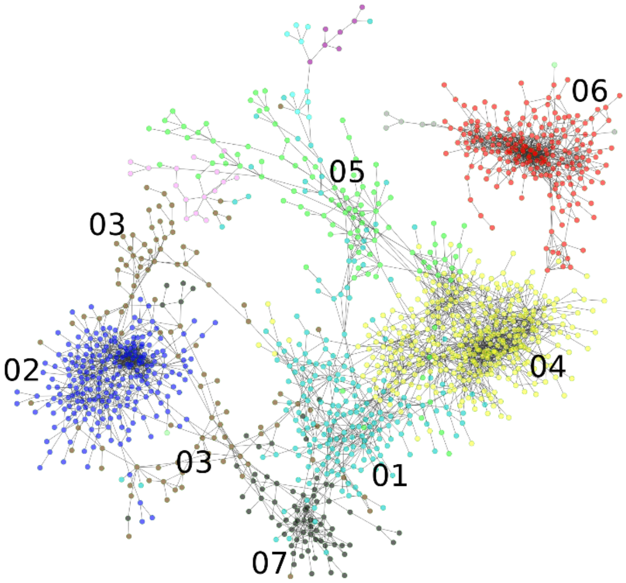


Figure 4. Weighted gene co-expression network of *Brachypodium* stress responsive genes. Major network modules are labeled by proximal numbers, which are identical to those listed in **Tables 1**, **2**, and **3**. Tight node grouping indicates mutually strong edges and therefore high adjacency. All adjacency values plotted are greater than 0.45.

This analysis resulted in a network that grouped 6,399 genes into 22 modules, the most strongly interconnected of which are shown in **Figure 4**. The expression profile of each module is shown in **Figure 5** as the average difference in RMA expression level between treatment and control arrays. The modular response of Brachypodium to abiotic stress was dominated by expression changes in response to the drought stress (**Figure 5**). Differential expression of modules in response to stress was determined by a requirement that an average expression profile must differ from that of the control by one RMA-normalized expression value at one time point under the given stress. Using this criterion, only one module was not responsive to drought stress (module 21; **Figure 5**, lower left). Nineteen of the 22 modules were either stress-specific in their response or responded to only one other stress in addition to drought stress. The remaining three modules are module 16, module 02, and module 07, which were all down-regulated in response to heat, high salinity, and drought stresses. No module was responsive to all four abiotic stresses. Lists of genes in each of the 22 modules may be found in **Supplemental File 1**.

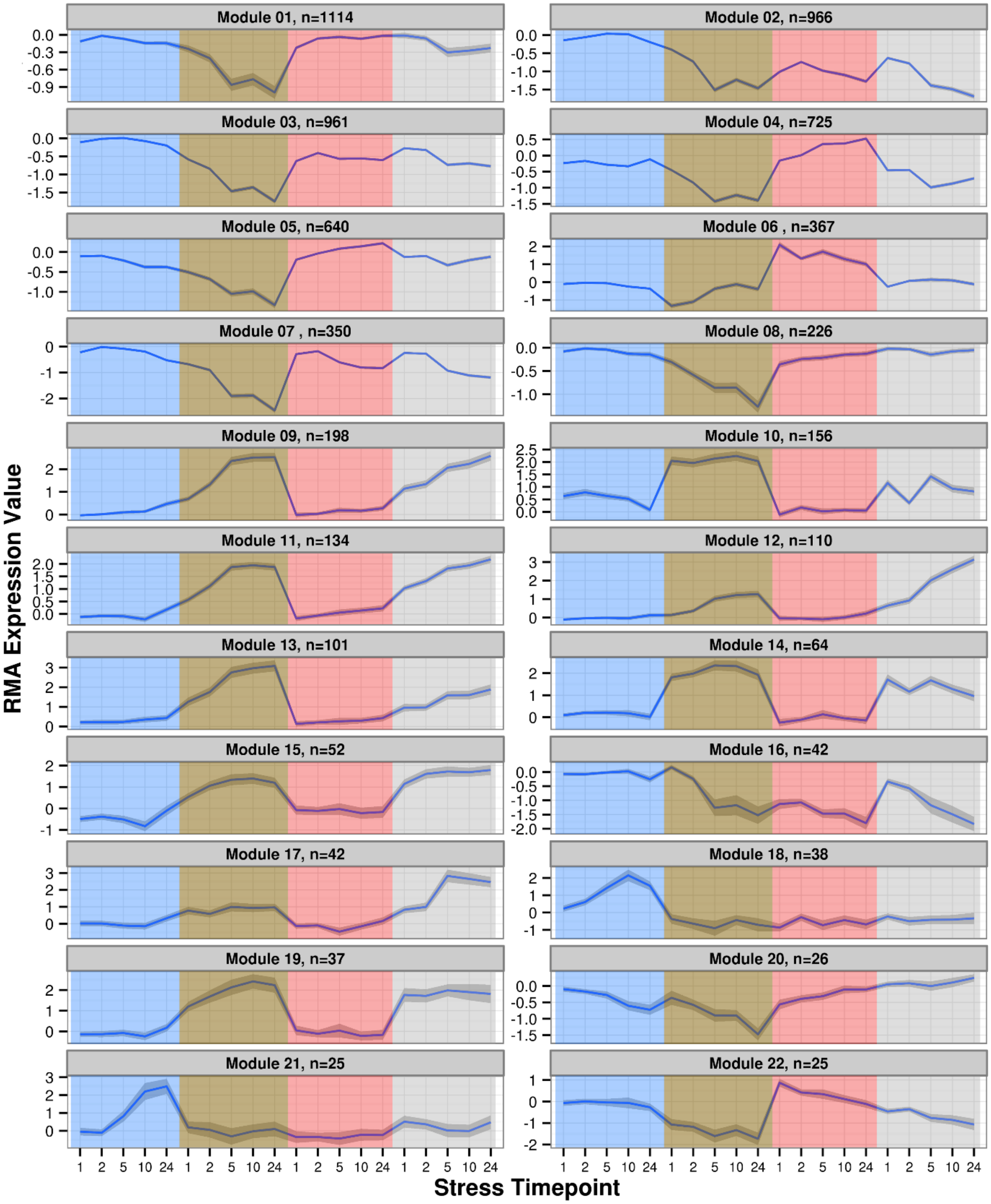


Figure 5. Expression profiles of modules as a function of time in each stress condition. Shaded area around lines indicates standard error. Values plotted are the average point-by-point RMA expression value differences between control and stress arrays for the member genes of the

module. N indicates the cardinality of the module in question. Color overlays indicate stress, from left to right: cold (blue), drought (brown), heat (red), and salt (grey).

2.2.3 Functional Annotation and Promoter Analysis

The combination of the functional annotations of the genes that comprise these modules with their expression profiles shed light on how the plant responds to abiotic stress conditions. Co-regulation is undoubtedly achieved through a combination of transcriptional and post-transcriptional regulation. The grouping of genes facilitated direct analysis of promoters to identify condition-specific over-represented cis-regulatory DNA elements. To assign functions to the modules, the module gene lists were analyzed using AgriGO (http://bioinfo.cau.edu.cn/agriGO/analysis.php) [26]. We also analyzed 500 nucleotides from the promoter regions of each of the genes in each module using the Element software package to identify over-represented DNA elements [27].

Table 2. Module membership and functional and regulatory enrichment.

Module	N	Undefined Genes	Unique GO terms	Total GO terms	Unique DNA Elements	Total DNA Elements
Module 01	1114	96	20	81	60	235
Module 02	966	70	59	75	299	441
Module 03	961	74	27	53	56	208
Module 04	725	39	55	101	323	504
Module 05	640	52	0	0	90	225
Module 06	367	18	11	13	107	151
Module 07	350	18	54	110	97	145
Module 08	226	22	0	0	5	24
Module 09	198	15	1	7	12	45
Module 10	156	6	0	15	190	354
Module 11	134	5	3	4	0	8
Module 12	110	2	0	0	32	69
Module 13	101	7	0	0	8	50
Module 14	64	4	0	0	9	37
Module 15	52	0	0	0	4	12
Module 16	42	1	1	2	3	17
Module 17	42	0	0	0	8	13
Module 18	38	2	4	25	1	15
Module 19	37	3	0	0	1	26
Module 20	26	4	0	0	0	0
Module 21	25	2	0	0	6	12
Module 22	25	1	0	0	1	1

The module-wise enrichment of GO terms and DNA sequences contained in promoters is shown in **Table 2**. There was a moderate correlation between the number of genes in the module with both the number of GO terms and with the number of DNA sequence elements found to be enriched within that module (Pearson's r: 0.616 and 0.755, respectively). This general correlation between module size and enrichment discovery is expected; however, there were exceptions to this general trend. For example, module 05 is ranked fifth in module size, with 640 member genes, but was not enriched for any GO terms (**Table 2**), although most (585) genes were associated with at least one GO term. Eleven modules were not enriched for any GO terms, and twelve were not uniquely enriched for any GO terms. The modules with no GO-term enrichment varied in size from the minimum size (N=25) to 640 members (module 05) (**Table 2**, column 'N'). Upon examination of the GO-terms enriched in each particular module, a pattern of enrichment was often apparent. A selection of the GO-terms enriched in each module, along with the relevant statistics, is shown in **Table 3**. AgriGO output for all 22 modules may be found in **Supplemental File 2**.

Table 3. Specific GO terms uniquely enriched in a selection of network modules.

Module	GO-term	Description	FDR
Module 01	GO:0004812	aminoacyl-tRNA synthetase activity	1.90E-05
	GO:0006418	tRNA aminoacylation for protein translation	8.80E-06
	GO:0006800	oxygen and reactive oxygen species metabolic process	0.022
	GO:0005525	GTP binding	0.039
	GO:0016875	ligase activity, forming carbon-oxygen bonds	1.90E-05
Module 02	GO:0007049	cell cycle	0.0059
	GO:0006260	DNA replication	3.30E-05
	GO:0034728	nucleosome organization	0.00045
	GO:0009832	plant-type cell wall biogenesis	0.00063
	GO:0000271	polysaccharide biosynthetic process	0.016
Module 04	GO:0003899	DNA-directed RNA polymerase activity	7.80E-07
	GO:0006281	DNA repair	0.00082
	GO:0033279	Ribosomal subunit	3.40E-13
	GO:0006364	rRNA processing	1.60E-09
	GO:0008026	ATP-dependent helicase activity	0.00091
Module 06	GO:0031072	heat shock protein binding	0.0012
	GO:0006457	protein folding	2.00E-21
	GO:0009408	response to heat	4.40E-19
	GO:0050896	response to stimulus	4.70E-04
	GO:0010035	response to inorganic substance	0.0043
Module 07	GO:0015979	Photosynthesis	3.20E-45
	GO:0033014	Tetrapyrrole biosynthetic process	1.90E-10
	GO:0006091	generation of precursor metabolites and energy	2.60E-21
	GO:0009765	photosynthesis, light harvesting	2.90E-18
	GO:0010114	response to red light	1.90E-06
Module 09	GO:0009415	response to water	0.0094
Module 11	GO:0009072	aromatic amino acid family metabolic process	0.0062
	GO:0022804	active transmembrane transporter activity	0.038
Module 18	GO:0006351	transcription, DNA-dependent	0.0018
	GO:0016070	RNA metabolic process	0.0076
	GO:0065007	biological regulation	0.0084
Module 16	GO:0016740	transferase activity	0.0088

Even in small modules with the minimum number of genes and no GO-term enrichment, we found over-representation of certain DNA sequences in member gene promoter sequences. Only module 20 was not enriched for any GO terms and had no over-represented DNA elements (**Table 2**). The over-representation of short regions of DNA sequence in the promoters of module member genes may provide insight into the transcriptional circuitry that mediates the regulation of the module. Twenty-one modules had at least one significantly over-represented DNA element (FDR-corrected p-value <0.01). Only two modules had no unique significantly over-represented DNA elements (**Table 2**, modules 11 and 20). Nine of the 22 modules had at least 32 unique elements over-represented in the promoters of their member genes (**Table 2**, column 'Unique DNA Elements'). Especially in conjunction with the functional annotation of modules via GO-term enrichment, the specific DNA elements which were uniquely enriched show how the transcriptomic responses of Brachypodium to abiotic stress compare to other plant systems (**Table 4**). In total, 1,312 elements of 5 to 8 nucleotides long were uniquely associated with specific modules. Element output pertaining to significant DNA motifs can be found in http://www.danforthcenter.org/hpriest/Supplemental_File_2.xlsx

Supplemental File *3*.

Table 4. Specific short DNA sequences found to be statistically enriched in the promoters of module member genes.

Module	DNA Element	Number of Hits	Number of Promoters	FDR
Module 01	TTAAAAA	346	267	4.94E-08
	TTTAAAA	301	197	1.71E-07
	CTCGTC	423	342	3.52E-05
	ACGTGGGC	139	120	6.03E-05
	CGGCC	380	299	4.80E-05
Module 02	CAACGGTC	57	48	3.79E-17
	AACGGCT	90	79	1.02E-09
	AGCCGTTG	47	39	2.43E-09
	CCAACGG	121	104	2.43E-08
	CAACGGC	115	98	5.38E-05
Module 04	AAACCCT	311	248	2.02E-69
	AGCCCAA	161	134	1.86E-14
	AGGCCCA	211	169	1.02E-28
	AAGCCCAT	57	50	2.57E-11
	GCCCAAC	115	100	1.86E-08
Module 05	ACAAAA	550	345	2.00E-05
	CAATA	617	368	7.05E-08
	ACAATA	197	151	4.04E-05
	ACAATAA	80	71	6.02E-06
	AATAA	1078	463	1.71E-05
Module 06	GAACCTTC	33	30	3.47E-15
	CTAGAAG	55	46	9.78E-11
	CTTCCAGA	28	26	3.98E-10
	AAGCTTC	61	40	1.01E-07
	GAAGCTTC	20	20	1.04E-06
Module 07	ACGTGGC	69	55	4.83E-12
	CCACGTC	59	53	1.39E-07
	GACGTGGC	25	21	5.88E-06
	CACGTGGC	26	20	1.27E-06
	CCTATC	92	81	1.12E-09
	GGGATA	83	78	7.11E-07
	AGATAA	126	105	0.00026
Module 09	ACGTAT	50	32	3.91E-05
	ACGTATA	23	14	1.14E-05
	ACACGTA	31	28	1.38E-06
	CACGTAC	36	28	1.29E-05
	CGTAA	118	83	0.000276
Module 10	CGATCG	47	35	0.00227
	CCGATCG	28	18	0.00049
	ATCGC	122	83	0.00424
Module 12	GTACGTA	27	13	6.08E-06
	GTACAC	41	36	1.44E-05
	ACGTACG	27	14	2.08E-05

2.2.4 Unknown Module Members

The lists of genes in modules were searched for genes which were identified as lacking useful descriptive annotations or as encoding proteins of unknown function. In all, 3,492 of 26,552 genes in the *Brachypodium* annotation version 1.2 were identified as lacking functional descriptions. In addition to those genes which are of interest due to the combination of their functional annotation and expression profile, genes without functional descriptions can be implicated in specific roles in abiotic stress, even if their function is unknown. The population of genes which are both unknown and members of modules are shown in **Table 2**.

2.2.5 Network Plasticity

Plasticity of gene regulatory circuits is an expected property of biological systems. There are multiple methods by which the expression relationship between a regulator gene and a target gene may change in response to varying conditions. The regulatory relationship between such gene pairs may change as a result of chromatin rearrangement or DNA methylation [71,72], both of which have been shown to be responsive to stress in plant species [73,74]. It is also conceivable that the abundance of the mRNA encoding a particular regulator could be detached from the target expression levels by protein modifications that alter either the activity or degradation rate of the protein in question [75,76]. The expectation that a transcription factor and target gene pair which interacts will generate correlated expression measurements may not reflect biological reality in all cases.

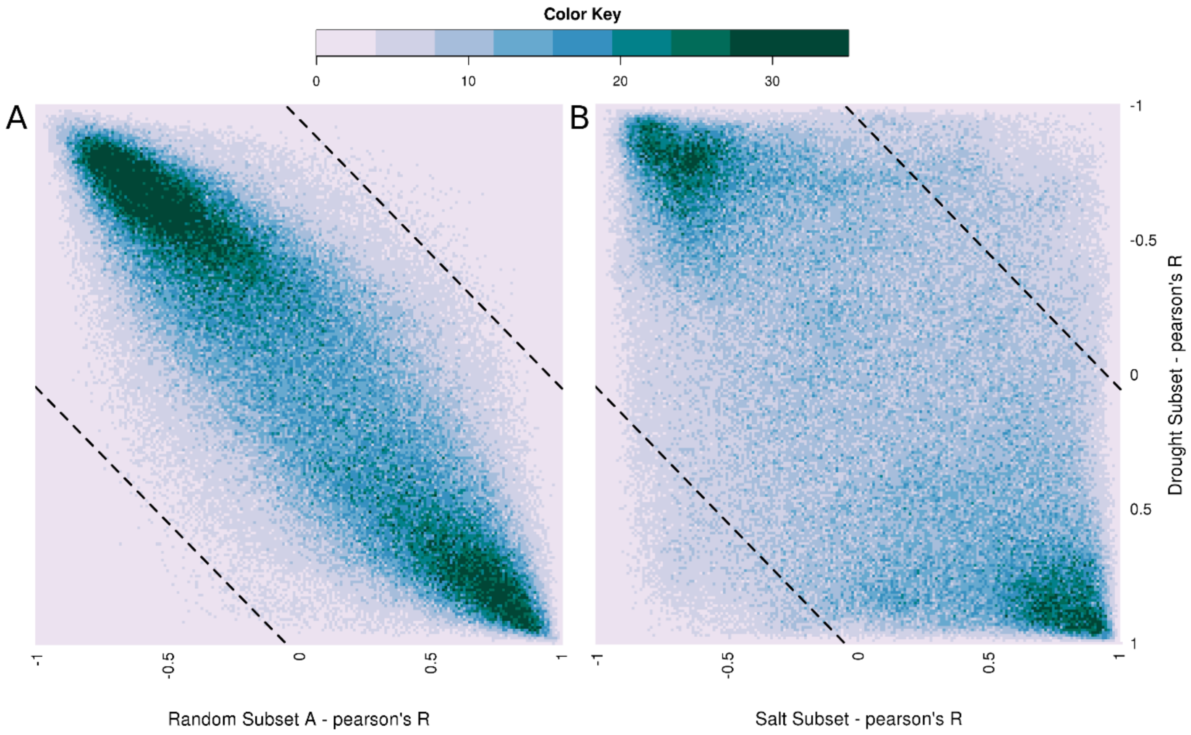


Figure 6 Scatterplot of transcription factor/target gene correlations. The x- and y-coordinates of any single pair of genes is determined by their correlation in the indicated subset. Colors are determined by the number of pairs that fell at a particular point according to the scale shown. Dashed lines indicate the minimum difference required before a TF-TG pair's correlations were considered significantly different between conditions. **A.** The correlations of TF-TG pairs in a random subset of data is compared against the correlations of those pairs in the drought assays. **B.** The correlations of TF-TG pairs in the salt stress and drought stress datasets are plotted. Large amounts of scatter are observed, in contrast to limited scatter in random samples, indicating that when compared across conditions, TF-TG correlations can be highly plastic.

Figure 6 shows heatmap-scatterplots of transcription factor/target gene (TF-TG) pairs in correlation space. TF-TG pairs are plotted according to their pairwise correlations in each of the shown conditions. Transcription factor/target gene pairs are defined as all possible pairings of genes differentially expressed in the two conditions of interest. Transcription factors are defined via a combination of sequence homology and InterProScan results (see Methods) [77]. The x-coordinate of a TF-TG pair is determined by the pairwise Pearson's correlation between that TF-TG pair in the indicated subset of stress data. The y-coordinate of that TF-TG pair is determined by the pairwise Pearson's correlation of that pair in the subset of stress data drawn from the drought experiment. The heatmap value is determined by the total number of TF-TG pairings with any particular combination of correlations. **Figure 6A** shows the distribution of pairwise TF-TG correlation changes between a random subset of the stress data and the subset of data drawn from the drought experiment, as an indication of what would be expected based on random changes of expression patterns. **Figure 6B** shows the distribution of pairwise TF-TG correlation changes between salt and drought stress data subsets.

In the salt-drought comparison, 146 TFs and 1910 non-TF genes were differentially expressed under both stress conditions. Based on the calculated threshold of Δr = 0.97 for the salt and drought comparison (see Methods), 27,916 of 276,950 TF-TG pairings (10.1%, **Table 5**) showed significant differential correlation across conditions, indicating possible plasticity in the relationship between the TF and TG of the pair (**Figure 6B**, top right and bottom left). The remaining 249,034 gene pairings showed less than significant changes in correlation across conditions. **Figure 6A** shows a representative distribution of correlation changes between gene pairs populated by a random permutation of the same data underlying **Figure 6B**. In distributions created by random permutation, an average of 1368.1 gene pairs per permutation were found to have significant changes in correlation based on the threshold of Δr = 0.97 for the same salt-drought comparison, corresponding to the targeted maximum FDR of 0.05 or less (**Table 5**). In all pairwise stress condition comparisons, between 0.9% and 24.9% of gene pairings were found to have potentially plastic relationships (salt/heat and salt/cold, respectively, **Table 5**).

Table 5. Putative network plasticity present between all pairwise conditional comparisons.

Stress A	Stress B	Gene Pairings	Plastic Pairs	Average False Positives	FDR	Δr cutoff
Drought	Salt	276,950	27,916 (10.1%)	1368.1	0.049	0.97
Drought	Cold	16,665	2,921 (17.5%)	144.9	0.049	0.96
Drought	Heat	70,434	4,890 (6.9%)	239.9	0.049	0.98
Salt	Heat	26,562	241 (0.9%)	11.9	0.049	1.35
Salt	Cold	8,132	2,027 (24.9%)	94.8	0.047	0.94
Heat	Cold	522	128 (24.5%)	6	0.047	0.88

2.2.6 Stress Responsive Modules in Brachypodium Transcriptional Circuitry

The motivations behind linking groups of genes to specific expression profiles in response to stress are multifold. First, modules represent regulatory relationships, indicating how *Brachypodium* reacts in a transcriptional and post-transcriptional manner to abiotic stresses. Second, the expression profiles themselves allow interrogation of the transcriptional regulatory circuitry that allows *Brachypodium* to achieve steady-state levels of stress-responsive transcripts at the appropriate time. This provides links between specific sequences present in the upstream regions of genes, key regulators (e.g. transcription factors), and traits of agricultural and economic interest.

Of all differentially expressed genes, 3,097 (32.6%) were not associated with a module. Different applications of stress, stress treatment severity, temporal distribution of sampling, and temporal density of sampling may enable association of many of these genes with these or other modules to more completely describe the stress response system of *Brachypodium*. Here, four abiotic stress treatments were used: heat, drought, high-salinity, and cold. We did not examine abiotic stresses such as high intensity light, UV, or chemical inducers of reactive oxygen species (ROS). With data on additional stresses, we will be able to associate more genes with over-arching modes of stress response.

2.2.7 Conserved Abiotic Stress Responses

Photosynthesis. Several sub-systems in plants are affected by multiple stresses. Photosynthetic activity (either capacity or efficiency) is known to be down-regulated or depressed upon heat stress [78], drought stress [79], salt stress [65], and cold stress [64]. One of the modules we identified, module 07 (**Figure 5,** top left), is comprised of 350 genes that are very strongly enriched for genes annotated with GO-categories related to photosynthesis, chlorophyll biosynthesis, light response and harvesting, and the chloroplast (**Table 3**, **http://*www***.danforthcenter.org/hpriest/Supplemental_File_1.xlsx

Supplemental File 2). For example, of the 143 genes in *Brachypodium* annotated with GO:0015979 'Photosynthesis', 50 are present in this module (a significant enrichment with FDR-corrected p-value of 3.2 x 10^{-45}). This module was down-regulated in drought, heat, and salt stresses (**Figure 5**). This indicates that under abiotic stress *Brachypodium* down-regulates photosynthesis as observed in several other plant systems [64,65,78,79]. As these genes associated with photosynthesis are affected by several stresses in a coordinated manner, these stresses likely modulate a common transcriptional circuit.

Eight genes in module 07 were found to be unannotated (see methods) – these loci were investigated further using the comprehensive Phytozome database (phytozome.net) [80]. This search revealed that these loci do not have functional annotations in *Brachypodium*, nor do their best homologs in other monocot species have functional annotations either. The co-expression of these genes with the other genes in module 07 indicates that they likely have some role in mediating either photosynthesis, or the regulatory response of photosynthesis-related genes to abiotic stresses in *Brachypodium*. The function of each of these loci must be elucidated by molecular and genetic analysis.

The ABRE (ACGT-containing abscisic acid response element) is a known *cis*-regulatory motif in *Arabidopsis thaliana* that contains an ACGT core and is responsive to drought [81]. This sequence was found in the promoter regions of many genes in the photosynthesis module (module 07), the water-response module (module 09, **Table 3**) and a transcription factor enriched module (module 10, **Table 3**, **http:**//***www***.danforthcenter.org/hpriest/Supplemental_File_1.xlsx

Supplemental File 2). Notably, even though the photosynthesis module and the signaling module (module 03) share highly similar expression profiles, this core sequence was not significantly enriched in the promoters of genes in the signaling module. The photosynthesis module is down-regulated under drought stress, whereas modules 09 and 10 are up-regulated under the same stress (**Figure 5**). Thirteen variations of the ABRE (including the ACGT core with differing flanking regions) were found in the photosynthesis module (**Table 4**, **http:**//***www***.danforthcenter.org/hpriest/Supplemental_File_1.xlsx

Supplemental File 2). Negative regulation of the photosynthesis module by the ABRE in response to drought stress was expected based on previous studies [82–84]. Forms of the ABRE were also over-represented in the promoters of genes in modules 11, 12, 13, 14, 15, and 19. These modules were not found to be over-represented for any GO-terms. However, these modules were up-regulated by both salt and drought stresses. The functional roles of these modules remain to be explored.

The photosynthesis module (**Figure 4, Table 3**) is strongly enriched for genes related to photosynthesis and was severely down-regulated in drought and moderately down-regulated in heat and salt stresses. These genes were not down-regulated in cold stress, but the overall depression of photosynthesis-related genes appears to be conserved in *Brachypodium* (**Figure 5**, top left). The relative strength of the stress conditions applied no doubt plays a role in the relative levels of regulation observed for this module.

Plant growth. Plant growth is severely affected by environmental conditions such as cold, high-salinity, drought, and heat [48,49]. Module 02 (**Figure 4**) is characterized by an expression profile similar to the photosynthesis module (module 07), though it shows larger negative changes in expression under both salt and heat stress treatments. Module 02 is enriched for genes annotated with GO-terms related to DNA replication, chromatin and nucleosome assembly, the cell cycle, and cell wall biogenesis (**Table 3**, **http://www**.danforthcenter.org/hpriest/Supplemental_File_1.xlsx

Supplemental File ***2***). The down-regulation of these genes suggests that an early response of *Brachypodium* to abiotic stresses is to suppress cell growth, DNA replication, and the cell cycle.

Similar to those genes in module 07, no functional annotation could be attributed to 77 loci in module 02, though they are differentially expressed in response to abiotic stress, and co-express with the rest of the genes of module 02. Given that these genes are co-expressed with the rest of the genes in module 02, it is likely that they play some role in the functions that are associated with their module, such as the cell cycle, DNA replication, or cell wall biogenesis. The specific functions of each of these genes must be described in follow up molecular and genetic experiments.

The Mitosis-Specific Activator (MSA) motif includes the core sequence ‘AACGG’ and is associated with G2/mitosis specific genes in *Arabidopsis* [85]. At*MYB3R4* has been shown to directly bind to this motif *in vitro* [85]. Module 02 is enriched for GO categories related to DNA replication, microtubule-based processes, chromatin, and nucleosome assembly. Thus, the 'cell-cycle' module is down-regulated under stress, indicating a suppression of these systems, which may result in a lengthened G2 phase and a slowed cell cycle. The promoters of the cell-cycle module are heavily enriched with the ‘AACGG’ core of the MSA motif, as well as its reverse complement (**Table 4**). Notably, the sequence ‘AACGG’ was found 907 times in 540 of the 966 gene promoters in this module (FDR-corrected p-value = 0.00043). Six distinct 8-nucleotide sequences containing this core were found 275 times (all six with FDR-corrected p-value <3.94 x 10^{-5}, **Table 4**). This core was also enriched in module 10; we observed this sequence 168 times in 95 of the 156 promoters (FDR-corrected p-value = 0.001,**http://*www***.danforthcenter.org/hpriest/Supplemental_File_2.xlsx

Supplemental File 3). Small plant stature and decreased yield are a major consequence of abiotic stress in plants [48,49]. A decrease in expression of genes activated by the MSA motif could conceivably result in a much slower or completely suspended cell cycle in the G2 phase. *Arabidopsis* plants deficient in TFs associated with the MSA showed pleiotropic dwarfism and other developmental and growth defects [85]. The putative ortholog of At*MYB3R4, Bradi2g31887*, is a member of the signaling module (module 03). The signaling module is also enriched for microtubule related GO-terms, as well as many signaling-related GO-terms. However, none of the unique significantly enriched DNA sequence elements present in the promoters of module 03 contain the MSA core nor is the MSA core itself enriched in gene promoters from this module (**Table 4, http://*www***.danforthcenter.org/hpriest/Supplemental_File_2.xlsx

Supplemental File 3). Elucidation of the relationship between the MSA and TFs such as that encoded by *Bradi2g31887* that may bind the MSA and suppression of the cell cycle by down-regulation of MSA-controlled genes will require further study.

Calcium-mediated stress response. Calcium receptors and calcium-binding proteins are important components of plant abiotic stress response. Calcium levels increase early in the cellular response to cold stress [86], and a link exists between calcium binding proteins and the cold-response CBF pathway in *Arabidopsis*. A model was recently proposed linking an increase in cellular Ca^{2+} levels with positive transcriptional control of CBF/DREB loci in *Arabidopsi*s [61]. Calcium levels also play a key role in drought and salt stress responses. At*CBL1* is an *Arabidopsis* calcium sensor that is up-regulated in response to salt, drought, and cold stresses [87]. Evidence suggests that calcium sensing plays a role in heat-stress response in monocot species as well [88–90].

Using homology to other model systems combined with annotation via InterProScan, 359 genes were associated with GO:0005509 (‘calcium ion binding’) or were associated with the phrase ‘calcium binding’. Expression data for these genes was hierarchically clustered and plotted in a heatmap (**Supplemental Figure 2**) that shows the expression of calcium ion binding genes in *Brachypodium* in response to the four assayed stresses. The expression levels of calcium ion binding loci were strongly affected by abiotic stress and were highly-correlated in drought and salt responses, although were independent in heat and cold stress responses. Principal component analysis of the expression data of the 359 genes annotated with GO:0005509 (**Supplemental Figure 3**) revealed that trends in expression of the 359 genes were highly similar to the trends in expression of differentially expressed genes overall. The first principal component was the strongest factor in later hours of drought and salt stress and explained 65.44% of the total variance of the expression data associated with the 359 putative calcium ion binding loci.

Of the 359 putative calcium ion binding loci, 88 genes were part of a module. This is significantly fewer than would be expected by chance alone (average expected overlap: 242 genes, Z-score -18.1). Sixteen of the 22 modules contained at least one putative calcium-binding locus. No module was enriched for GO:0005509 ('calcium ion binding'). The large distribution of calcium responses to abiotic stress (**Supplemental Figure 2**) indicate that there are multiple regulatory pathways that trigger calcium ion binding protein expression and that these loci play a role in mediating the response of *Brachypodium* to the four assayed stresses. Further, their significant under-representation among modular loci suggests that the response of individual differentially expressed calcium loci does not conform to the major modes of stress response. The regulatory circuits that control calcium ion binding loci appear to be specific to these individual genes. Prior studies provided evidence that calcium ion levels, calcium ion binding protein levels, and abiotic stress responses are linked in multiple plant systems [61,87,90]. Our analysis confirms that calcium ion sensing and calcium ion binding loci are responsive to abiotic stress in *Brachypodium*. We found no evidence of a centralized calcium response system.

Novel and uncharacterized modules. Module 05 is down-regulated under drought stress but not differentially expressed under any of the other three stresses. Module 05 was not enriched for any GO terms (**Table 2**). Of the 640 genes in the module, 585 genes were annotated with at least one GO-term. The promoter regions of the genes in this module were enriched for 225 specific conserved motifs; of these, 90 are uniquely enriched in module 05 (**Table 2**). These include the core CAATA (FDR-corrected p-value $7.05x10^{-8}$) and the variant ACAAAA (FDR-corrected p-value $2x10^{-5}$). The PlantCARE [91] database lists the core CAATA as part of an Auxin Response Element (ARE) in *Glycine max.*

Like module 05, module 08 is down-regulated only in drought. This module has 226 member genes and is not enriched for any GO terms. Twenty-four DNA sequence motifs were significantly enriched in promoters of module 08. Uniquely significant motifs included TCCTTCA, CCCGAC, and CCGAAA. These motifs are similar to the CRT/DRE DNA TF-binding site, RCCGAC [92,93]. Conserved *cis*-acting elements similar to those found in the promoters of modules 05 and 08 have been observed in other species, lending weight to the hypothesis that these DNA sequences could be responsible for driving the module-wise expression profiles observed here. No enriched functional terms could be associated with modules 05 and 08. An extended examination of gene expression responses to abiotic stress – especially stretching into the days after stress onset – may reveal the functional roles these modules play.

2.3 Discussion

This study provides insight into the regulatory responses of *Brachypodium* to four abiotic stresses. Application of the *Brachypodium* genome-scanning tiling array resulted in deep profiling of the transcriptional response to abiotic stress. The data and analysis provided here will be an excellent resource for researchers utilizing *Brachypodium* as a model system, as will the web-based resources provided for community use.

2.3.1 Conserved Modular Responses

Previous studies in rice observed a high overlap between gene sets differentially expressed in response to drought and high salinity stresses [6]. Our work captures a similar response in Brachypodium, with roughly 75% of the genes differentially expressed in response to salt also differentially expressed in response to drought. Similarities in overall pattern and variance of the responses to drought and high-salinity are also seen in the Principal Component Analysis (PCA).

Many systematic responses to abiotic stress in Brachypodium could be characterized on the modular level – these responses are coordinated in independent stresses. This is reflected in the very strong enrichment of photosynthesis-related genes in module 07 (**Table 3**), and the expression pattern of the same module in response to drought, heat, and high-salinity stress (**Figure 5**). The well-characterized behavior of photosynthesis systems in response to stresses [19,20,35,36], combined with the distinct co-expression profile of module 07 lends further weight to the hypothesis that this response is a coherent systematic response mediated by an underlying gene regulatory network. Strong similarity between regulatory motifs (**Table 4**) found to be enriched in promoters of stress-responsive genes in Brachypodium to those identified in stress experiments in Arabidopsis [38,42] suggests that similar circuits are present in Brachypodium. Similar coherency of response was observed for genes related to the cell cycle, as well as conservation of upstream regulatory sequences related to mitosis.

In contrast to the clear coherency of transcriptional regulation of the photosynthetic system, no such coherency was observed for genes related to calcium signaling and binding. Calcium ion binding related loci were sequestered out of modules at a highly significant level (Z-score = -18.1, two-tailed p-value < 1e-6), which indicates that unlike more coherently regulated systems, calcium ion binding does not co-express strongly with other genes. Taken in conjunction with the knowledge that calcium-ion binding loci are important for plant abiotic stress response [16], this indicates that the transcript-level expression of these loci simply is not in line with the major modes of plant stress response captured in these experiments.

2.3.2 Network Plasticity

Analysis of differential correlations for transcription factor/target gene pairs in various conditions revealed a high degree of plasticity in these relationships. The proportion of potentially plastic relationships varied greatly depending on the conditions compared. Neither the conditional comparison with the lowest ratio of potentially plastic gene pair relationships (salt/heat, 241 plastic TF-TG pairs, **Table 5**) nor the comparison with the highest ratio of potentially plastic relationships (salt/cold, 2,027 plastic TF-TG pairs, **Table 5**) were the comparison with the most extreme number of total possible pairings. Of particular interest is the great diversity of differential correlations between salt and drought stresses. There are a large segment of gene pairs that experience very large changes in correlation. More than 11,000 genes pairings had large negative correlations under drought stress and very large positive correlations under salt stress (top right, **Figure 6B**). Conversely, more than 16,000 gene pairings had large positive correlations under drought stress and large negative correlations under salt stress (bottom left, **Figure 6B**). Comparisons between the differential correlations observed between salt and drought stresses and the differential correlations observed between random subsets of the stress data indicate that the differential correlations between salt and drought stresses are unlikely to arise by chance (**Figure 6A**).

The basic underlying assumption of gene co-expression network analysis is that two genes, when co-expressed, can be expected to be reliably co-expressed if there is a biological relationship between them. The stronger the biological relationship between two genes – either due to genuine co-regulation or from necessary co-expression borne of functional relatedness, the higher the correlation in expression between the two genes. The relationships between transcription factor/target gene pairs across conditions are plastic due to dependence on DNA methylation and chromatin modification status, among many other factors. This highlights the importance of inclusion of epigenomic data in any large genomic discovery endeavor.

Because of the possible relationship between TF loci and their target genes, we queried the module membership of the TF loci population, to determine if they were preferentially included or excluded from modules. Similar to the exclusion of calcium ion binding loci from modules, the exclusion of TF loci from modules would indicate that they are more selectively regulated in response to abiotic stress than the loci which are identified to be module members. Of 600 TF loci which are differentially expressed in response to stress, 369 are members of modules. This is significantly fewer TFs than would be expected by chance alone (determined by permutation test, 404.5 loci expected, Z=-3.195, two-tailed p-value=0.0014). As modules are built on co-expression across many conditions, and it appears the gene co-expression correlations may be plastic, the expectation that TF-TG relationships are consistent across conditions may be incorrect, and the de-enrichment of TFs in modules may reflect that.

In addition to the sequestration of TFs out of modules – which may reflect the plasticity of their relationships to modular genes – genes which are distinctly lacking plastic relationships are of great interest. On the hypothesis that gene co-expression plasticity stems from changes in the underlying biochemical relationship between loci, genes which lack plastic relationships may lack the requisite biochemical changes in regulatory relationships, and may have stable regulatory circuits. Of the 2,752 genes which were considered in the plasticity analysis, 220 genes never showed any plastic relationships to any TF (7.9%). Put another way – the correlation changes across conditions between these genes and the TFs to which they were correlated was always below the significance threshold. Of these 220 genes, 29 were found to be un-annotated. The list of genes which had no plastic relationships also included *Bradi1g42630* annotated as a phosphofructokinse, a locus down-regulated in drought, salt and heat stress, which was a member of module 02. This gene was highly homologous to *AT1G76550*, an *Arabidopsis* phosphofructokinase. A member of this family in *Arabidopsis* was identified as one of a group of genes which influence plant growth and biomass [94].

A second non-plastic gene is *Bradi5g11640*, which is differentially expressed in response to drought and heat stresses. This gene is highly homologous to *AT1G65960* a glutamate decarboxylase which was found to have its enzymatic activity increase in response to treatment by calcium and calmodulin in combination, indicating that the *Arabidopsis* locus encodes a calmodulin binding protein [95]. The specific role of this locus in *Brachypodium* remains to be elucidated by further molecular experiments.

Sources of gene co-expression plasticity can stem from either the regulator or the target locus. Loci which have particularly stable relationships may represent a group of loci which remain highly accessible to the transcriptional machinery during the four assayed stresses. While this group of 220 genes may be hypothesized to be a ‘core’ group of stress reactive genes, these genes were not enriched for any particular GO term or category.

Based on the dataset used here, we cannot assign cause to the large changes in expression correlation across conditions. It is clear that a full understanding of the abiotic stress response of *Brachypodium* requires epigenomic analysis. With increasing throughput and decreasing costs, full integration of multi-type sequence data waits only on development of novel bioinformatic methods that can take full advantage of rich datasets. The high degree of plasticity observed in the stress response of *Brachypodium* also has implications for whole-genome gene co-expression network reconstruction. Current state-of-the-art software packages, such as WGCNA [96], may be made even more powerful by accounting for the changing relationship between gene pairs across conditions in meta-data enhanced expression datasets. Adopting a ‘regulator-target’ dichotomous view of genes – as is common in applications designed for smaller networks – may further improve large network reconstruction efforts.

Weighted gene co-expression analysis of the *Brachypodium* transcriptome under normal growth and four abiotic stress conditions identified 22 modules of genes. Over-expression, knock-down, and knock-out experiments will elucidate the roles of these genes in abiotic stress responses and may guide genetic approaches that confer stress tolerance in economically important grasses. This research provides insight into how this model crop system responds to abiotic stresses. Homology between *Brachypodium* and agricultural target species will allow the identification of stress-responsive target genes in cereal and biofuel feedstock crops, enabling improved stress tolerance in plants critical to serving the needs of society.

We have identified numerous potential transcription factor binding site sequences that are associated with specific expression profiles under abiotic stresses. In addition to correlating these motifs to specific gene expression profiles, we have linked these DNA sequence motifs to specific endogenous plant systems. These candidate *cis*-regulatory sequences may represent key components of the transcriptional circuitries that define the plant's gene regulatory networks. Systems and synthetic biology approaches may take advantage of these circuits to place genes of interest under the control of existing stress response pathways to achieve desirable phenotypes of stress tolerance in agriculturally or economically important crops.

2.3.3 Web Resources

All microarray datasets are accessible through the Brachypodium web genome browser (http://jbrowse.brachypodium.org). The module membership lists, AgriGo GO-enrichment analysis output, and Element promoter content analysis output may be found as supplemental files and are available for download on the Brachypodium.org FTP website (ftp://brachypodium.org/brachypodium.org/Stress/). All individual gene RMA expression stress response profiles for each assayed stress condition may be viewed at the Mockler Lab's plant stress response web portal (http://stress.mocklerlab.org/).

2.4 Conclusions

The results achieved here represent an excellent characterization of the abiotic stress response of *Brachypodium distachyon* to high soil salinity, high temperature, low temperature, and drought. However, the results shown in section **2.2.5** and **Figure 6** represent a key failing of the analysis presented here. The application of WGCNA to the set of all microarrays essentially ignores all those expression similarities which occur only in certain subsets of the dataset. The analysis above identifies modules which are responsive to stress. Most of the modules are responsive to more than one stress, and a small minority are stress specific. This is an artifact of the analysis design – gene pairs which are co-expressed only in one abiotic stress condition will not have a strong enough similarity when their expression patterns across all four abiotic stress conditions are analyzed as a whole.

The most proper approach to a dataset such as this would be to identify stress-specific gene co-expression networks, to compare those networks to identify those gene relationships that change in a significant manner. In addition to that analysis, the analysis above is also necessary, but also subnetworks which allow for all possible combinations of two and three abiotic stresses in conjunction (i.e., high salinity and low-temperature together, high salinity, drought, and heat, but not chilling, etc.). A rigorous method for network comparison, and an illustration of the edge sets identified by such an approach, is the topic of the next chapter.

2.5 Methods

2.5.1 Experimental Growth Conditions and Tissue Sampling

Brachypodium distachyon control plants were grown at 22 °C with 16 hours light and 8 hours dark in a controlled environment growth room. Abiotic stress conditions included cold, heat, salt, and drought. All treatments were conducted with a light intensity of 200 µmol photons $m^{-2}s^{-1}$. For the heat experiments, *Brachypodium* plants were placed in a Conviron PGR 15 growth chamber at 42 °C. Cold treatments were conducted in a walk-in cold room maintained at 4 °C. Salt stress (soil saturation with 500 mM NaCl) and drought (simulated by removing plants from soil and placing them on paper towels to desiccate) treatments were conducted under the same light and temperature as the control samples. Three-week-old *Brachypodium* plants were placed under the respective conditions two hours after dawn (10 a.m.). Leaves and stems (total above ground tissues) from individual plants were collected at 1, 2, 5, 10, and 24 hours after exposure to the abiotic stress.

2.5.2 RNA Preparation, Labeled cDNA Synthesis, and Microarray Hybridization

Leaf tissues were pulverized in liquid nitrogen, total cellular RNA was extracted using the RNA Plant reagent (Invitrogen), and RNA was treated with RNase-free DNase essentially as described in [97]. DNase-treated RNA integrity analysis, preparation of labeled target cDNA from *Brachypodium* leaf total RNA, Affymetrix microarray hybridizations, chip scanning, quality control, image processing, and data extraction were performed essentially as described in [98]. One array – heat-stress hour 5 replicate 'C' – did not pass quality control and was discarded.

2.5.3 Mapping of Probes

Probes on the Affymetrix BradiAR1b520742 array were mapped to the Bd21 v1.0 assembly using the Burrows-Wheeler Aligner (BWA) [99]. The Bd21 Brachypodium Array contains 6,503,526 non-control probes. Of these, 99.81% (6,491,341 probes) map to a single location in the genome. Most of the probes (6,491,341) match their target sequences unambiguously with no mismatches in alignment. Only 12,183 probes align with mismatches. All probe sequences represented on the array are entirely distinct from each other. For the probe-set level analysis, probes were associated with annotated genic features. Probes that associated with a single gene's exonic features were collected into strand-specific probe-sets. Only those probe sets associated with the forward strand of a target gene were retained for analysis in differential expression or network prediction. If a probe was associated with exonic features of two genes (if two genes overlap, for instance), that probe was not assigned to any probe set. If a probe was associated with both intronic and exonic features (if a gene has multiple transcripts, or a probe spanned an exon/intron boundary), the probe was not assigned to a probe set. In the 47,960 genic probe sets, each gene was detected by, on average, 31.5 probes. The median number of probes per set was 22.

2.5.4 Microarray Data Analysis

Probeset level expression values were obtained utilizing the Robust Multi-array Average [70] technique via the Affymetrix Power Tools (APT) software package (http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx). Probe set summarization and expression estimates for each gene were conducted using the apt-probeset-summarize tool (version 1.15.0) from Affymetrix. Data manipulations were performed using Perl scripts. From the resulting signal intensities, differentially expressed genes were calculated using the Significance Analysis of Microarrays (SAM) [100] R package in conjunction with Microsoft Excel.

SAM uses permutations of repeated measurements to estimate the percentage of genes that are identified by chance, representing the false discovery rate. SAM was run with default settings, using 100 permutations, using the ‘two class unpaired’ response type. The S_0 factor was estimated automatically and no fold-change cutoff was applied at the time of differential expression calling. The Delta value was selected such that the median false-discovery rate was below 0.01. In every case, control and stress RMA expression values were compared in a pairwise fashion within a single stress and time point combination.

2.5.5 Heatmap and Principal Component Analysis

Heatmap and Principal Component Analysis (PCA) analyses were conducted in R. RMA expression differences between the average expression value per stress time point per treatment were set to saturate at a difference of 4 RMA (such that the maximum value reported in the heatmap was +/- 4 RMA). These expression differences were graphed using the 'heatmap.2' function of the gplots package of R. For principal component analysis, the average RMA expression value of each stress time-point, without the above saturation, was used as input for the 'PCA' function of the R package 'factominer' (http://factominer.free.fr/) [101].

2.5.6 GO Analysis and Transcription Factor Annotation

Over-represented GO terms were identified using the AgriGO: GO analysis toolkit (http://bioinfo.cau.edu.cn/agriGO/) [102]. Analysis was done by comparing the number of GO terms in the test sample to the number of GO terms within a background reference. Over-represented GO terms had a FDR corrected *P*-value of less than 0.05 and more than 5 mapping entries with a particular GO term. GO-terms were assigned to genes based first on InterProScan [77] results for the entire predicted proteome of the *Brachypodium distachyon* MIPS version 1.2 annotation [103]. Approximately 40% of genes did not have any GO-terms associated with them. Gene products from this set that had high-quality BLASTP matches to *Arabidopsis thaliana* gene products were assigned the same set of GO terms that their *Arabidopsis* homolog possessed. The list of putative *Brachypodium* transcription factors was obtained from gene annotation queries and BLASTP comparisons to rice (*Oryza sativa*) transcription factors obtained from Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de/v3.0/) [104].

2.5.7 Network Analysis

Normalized RMA expression values for 9,496 differentially expressed genes were loaded into the R package WGCNA [96]. An adjacency matrix was calculated using B=23. Distance metrics between profiles were calculated using the TOMdist function using an un-signed TOM type. Hierarchical tree solution was calculated using the flashClust [105] function with the ‘method’ option set to ‘average’. Modules were called using the moduleNumber function, cutHeight=0.91, and minimum module size was set to 25. Module colors were set using labels2colors. These modules were merged, using mergeCloseModules, a cut height of 0.1, iteration set to ‘true’, and enabling re-labeling. Final module colors and numbers were set as a result of this merging. Modules were exported for visualization in Cytoscape [106] using the “exportNetworkToCytoscape” function in the WGCNA R package and an adjacency threshold of 0.35. Once imported to Cytoscape, edges were filtered for a minimum value of 0.45, and the final network layout was obtained using the “Force Directed” in-built Cytoscape layout method. Cytoscape-layout and edge filtering caused some modules to not be connected by edges. These were not included in final Cytoscape layout; however, their mutual connectivities in the adjacency matrix served to allow WGCNA to call them as modules so they were analyzed as such for AgriGO-mediated GO enrichment and for Element-mediated promoter analysis. Only those modules that were graphed in Cytoscape as being interconnected with edges above the 0.45 cutoff were included in the final figures.

2.5.8 Promoter Analysis

Genes were grouped based on module membership. Based on the MIPS version 1.2 *Brachypodium distachyon* annotation, the 500 nucleotides directly upstream of each gene was extracted from the *Brachypodium* genome. The promoters for the genes in each module were analyzed on a module-by-module basis using Element [21]. The set of all predicted promoters in the genome were analyzed using the ‘bground’ command using all possible 5 to 8 nucleotide sequences as the set for analysis. This formed the set of background motif occurrence statistics against which module groupings of promoters were compared. Motif occurrences in module sets of genes were then compared against the background set. Motifs shorter than 5 nucleotides in length are expected to fall into one of two categories – background false-discoveries or true-positives that will be contained within larger, also significant motifs. Transcription factor binding sites longer than 8 nucleotides in length are expected to either overlap or be multi-partite motifs, both of which will generate significant sub-motifs in this analysis. In some cases, for specific examples, membership lists from two modules were combined for analysis by Element. Element was run using default cutoffs for significance (FDR<0.01), on 16 processors ('-t 16').

2.5.9 Network Plasticity Analysis

Network plasticity was determined by comparing the correlation of gene pairs between conditions. Between two conditions, every gene that was called by SAM as being differentially expressed in both conditions was segregated into one of two groups – the TF group or the non-TF group. Putative *Brachypodium* transcription factors were identified as described above. All pairwise Pearson's correlation values were calculated between groups in each of the conditions. This yielded two correlation values for each gene pair – one value corresponding to each condition. The order of the values of each gene expression profile across all assayed stress conditions was then randomly shuffled via the Fisher-Yates Shuffle procedure [107] creating 7,200 random permutations of the data. In each permutation, two subsets of equal size (N=15) were selected. Each permutation therefore was a random permutation of a gene's total expression data profile from which two independent samples of size N=15 were selected. The pairwise Pearson's correlations between all TF-TG pairs were calculated in each permutation. In order to determine significance of correlation change across conditions, a cutoff was chosen such that the average number of genes pairs that had correlation changes exceeding that cutoff in each random permutation (average number of false discoveries per permutations) was an appropriately small ratio of the number of gene pairs that had correlation changes exceeding that threshold in the true dataset (number of positives). This process is similar to SAM [100]. In all comparisons, the threshold was chosen such that the FDR was less than or equal to 0.05.

2.5.10 Undefined Module Member Genes

In order to identify genes which could be associated with a role in abiotic stress response by module membership, but could not have a predicted function attached to them, the entirety of the *Brachypodium* proteome was aligned against the Phytozome annotations for *Sorghum bicolor* [108], *Glycine max* [109], *Arabidopsis thaliana* [110], *Zea mays* [111], *Setaria italica* [112], and *Oryza sativa* [113]. Proteins which aligned with 70% identity over 70% or more of their total length, to a gene in one of the target species were associated with the functional annotation of the target gene. Of 26,552 *Brachypodium* proteins, 15,480 (58.3%) aligned to at least one target gene in at least one target species. 11,072 genes (41.7%) did not align to any target genes in any target species. Of those genes that aligned, 1,313 were associated only with annotations such as “expressed”, “putative protein”, “protein of unknown function”, or similar, and never with more functionally-informative annotations. These were identified as undefined loci. In order to supplement these associations, InterProScan [77] annotations were included. Genes which did not have an informative InterProScan result, and did not align to a target species, or, did not have an informative annotation if they did align, were identified as undefined loci. Therefore, the only information we could reliably attach to these loci were their expression profile and the set of genes with which they co-express.

2.5.11 Accession Number

The raw data is available at the Plant Expression Database (www.plexdb.org) under PLEXDB accession number ‘BD2’.

ACKNOWLEDGEMENTS

We are grateful to Anne-Marie Girard and Caprice Rosato for the qualitative assessment of RNA, preparation of labeled targets, and executing array hybridizations and scanning. We thank the

Donald Danforth Plant Science Center's Biocomputing Core for computational support. We thank Malia Gehan for helpful suggestions on the manuscript.

dGCNA: edge-wise differential gene co-expression network analysis

3.1 Introduction

The advent of High-Throughput Sequencing has made the generation of large expression datasets a commoditized experimental assay. Economies of scale and improvements in sequencing technology drive costs down and data yields up. As costs decrease and yields increase, the complexity and scale of gene expression experiments expand. This great expansion in data scale and complexity presents two challenges. The first challenge is that it is difficult to extract biological meaning from large, many-faceted expression experiments. The result sets of pairwise differential expression tests or analysis of variance approaches rapidly become untractable and unmanageable. This challenge is overcome by the application of novel computational methods. This solution creates the second challenge, in that the individuals who are most suited to conducting complex computational analyses are often not suited to teasing apart biological meaning and identifying the best course of action in terms of candidate selection and experimentation.

3.1.1 Gene Co-Expression Network Analysis

In terms of transcriptome-scale network-based gene expression analyses, the application of gene co-expression network (GCN) analysis has become widely utilized. On the most basic level, a network is a collection of network “nodes” (in this case, genes) and network “edges” (pairwise, gene-gene relationships). An individual edge connects two nodes, and the collection of all nodes and edges make up the network as a whole. In a gene co-expression network, a node has an associated expression profile – this profile is the set of all gene expression values obtained in an expression experiment. In short, gene-gene expression profile similarities are determined by a similarity metric (such as the Pearson Correlation Coefficient, PCC, [114]). These similarities are passed through a transform ("adjacency function") to determine a binary adjacency matrix, in which edges that are related have entries of "1", and unrelated edges have entries of "0". The relationships described in the adjacency matrix can then be clustered, to generate sets of interrelated nodes, which have been shown to have functional significance.

Perhaps the most widely utilized framework for network analysis of large scale expression datasets is the Weighted Gene Co-expression Network Analysis (WGCNA, [96]) R package. This schema introduces a great deal of granularity into the final set of node relationships; the "weighted" aspect of the framework allows the final relationships to be decimal values on the interval [0,1], or [-1,1] depending on user preferences.

3.1.2 Network Comparison and Elasticity

Gene co-expression networks are a powerful tool to describe and characterize gene expression trends in large, complex datasets. A common initial conception of GCNs is that the inclusion of more data will increase the breadth and granularity of the GCN – resulting in a network describing the behavior of more genes and identifying smaller clusters of related genes. This turns out not to be the case. Feltus et al., showed that as more and more data was included in a network, fewer and fewer genes were reliably related to one another, and clusters became more globular and featureless [115]. Their work discovered instead that it was more practical to group expression assays that were themselves related, and to build many small networks rather than one large network. In recent work in *Brachypodium*, a large amount of elasticity was found to exist in the relationships between genes under varying abiotic stresses [69]. The implication is that, under varying conditions, gene co-expression networks are remodeled as the regulatory landscape of the underlying biological system changes to meet the needs of the organism.

These analyses point directly to a short-coming of the current schema of GCN analysis techniques, which was partially approached by Feltus et al.; as more datasets are added to a GCN analysis, only those pairwise relationships which are stable across all datasets are identified. The elasticity of gene-gene expression relationships all but guarantees that with enough data, no gene-gene relationship will hold across all observable perturbations of a biological system.

The comparison of GCNs has been approached multiple times. The DNA R package, was created to identify differential connectivity, edge strength, and structure. However, it is limited to small networks (e.g., 20-400 genes), and so addresses a different problem set [116]. The algorithm mlDNA [117] approaches a similar scale of problem as dGCNA. However, it is based on the identification of previously genes known to be responsive to the particular perturbation. This is not necessarily a problem, as in well-studied organisms, these genes will be known. However, dGCNA is targeted at identifying novel signal without prior information. The DINA algorithm is implemented in a web platform and is targeted as pathway-size gene sets. While clearly effective, it relies both on prior knowledge and a small gene set [118]. The original authors of WGCNA also put forth a strategy for differential network analysis, relying on differences in the whole-network connectivity measure for a gene between individual networks. This final method identifies genes which undergo large changes in connectivity. Indeed, as we show below these genes are enriched for information, however, this method does not identify statistically significant changes in the edges of genes which do not have very high overall connectivity, is not implemented in a software package or program, and is not immediately generalizable to any new comparison [119]. Indeed, this method is defined as a loose framework for differential network analysis, rather than an algorithm or software package proper. The most similar method of direct network comparison is that of DiffCoEx [120], to which we directly compare our results.

It is exceedingly doubtful that a single static GCN will ever describe the expression relationships between an organism's genes in all observed conditions. Instead, GCNs describe the expression landscape of – for example – a particular cell type, an environmental condition, or developmental stage. It will therefore be important to not only construct GCNs describing a discrete biological state, but to compare networks of various states to observe how the gene-gene relationships are remodeled to meet the needs of the organism.

To that end we have developed an algorithm, dGCNA, which directly compares two GCNs derived from closely related datasets. Implemented in Java, the algorithm greatly eases the difficulty of comparing a large, complex structure such as a GCN, and identifies statistically significant adjacency differentials in the edges and nodes between the GCNs of interest. Comparing two datasets via their emergent GCNs also allows the comparison of datasets of unequal size. We show the results of our algorithm on a published dataset of circadian expression derived from *Arabidopsis thaliana*, and illustrate the biological meaning derived from identifying co-expression elasticity within GCNs. We further demonstrate the fine-grained insight which is gained by comparing individual, conditional networks rather than analyzing them as a single whole. Finally, we compare our method to existing packages or algorithms with the stated goal of network comparison.

The source code of dGCNA is available for download at https://github.com/hdpriest/dGCNA.

3.1.3 Author Contributions

Todd Mockler directed the research focus to the topic of topological changes within gene co-expression networks, and contributed to the manuscript. Henry Priest conceived, designed, and developed the dGCNA algorithm and program, conducted analysis and wrote the manuscript. At the time of this writing, this material has not been published. Written permission for use of this material has been obtained from Todd Mockler.

3.2 Algorithm & Methods

3.2.1 Algorithm & Implementation

The dGCNA algorithm consists of two main parts – determination of appropriate parameters for the adjacency function, and application of those parameters in the generation of a differential GCN. Gene inclusion and exclusion is left to the user's criteria and preferences. Both datasets utilized for comparison must comprise the same overall set of genes. The number of observations in each dataset need not be identical. The software manual includes a tutorial, description of proper format for input expression data, advice for gene inclusion/exclusion, and selection of observation sets can be found in **http://www**.danforthcenter.org/hpriest/Supplemental_File_3.xlsx

Supplemental File 4.

Terminology

The purpose of the algorithm described in this manuscript is to identify statistically significant pairwise differential gene co-expression within GCNs. This is achieved by applying the method of adjacency transformation to a differential similarity matrix, rather than a similarity matrix. The terms 'elasticity' and 'differential adjacency' refer to two discrete entities. Each individual edge has some associated differential adjacency value, on the interval [-2,2]. However, these values are not directly representative of the resultant elasticity of the edge in question. Elasticity describes the property of an edge to become either more or less strong in network 2 relative to network 1. Negative elasticity refers to an edge which decreases in absolute value in network 2 relative to network 1. Positive elasticity refers to an edge which increases in absolute value in network 2 relative to network 1. An edge may have a positive differential adjacency, but negative elasticity (for example, a change from -0.9 in network 1 to -0.1 in network 2, a differential adjacency of 0.8), or a negative differential adjacency but positive elasticity (-0.1 to -0.9, the converse of above). Differential adjacency refers to a mathematical value; elasticity refers to a network edge-wise property. Regardless, the product of a comparison between two GCNs is a set of node-node interactions, with all edge values lying on the interval [-2,2] which we term a differential-Gene Co-expression Network, or dGCN.

Algorithm Process

The algorithm proceeds in a way similar to a classic GCN analysis. First, the datasets are analyzed, to provide the user with information on how best to select parameters for the adjacency function. Second, the datasets are permuted and iteratively compared, to estimate the statistical significance of any particular differential similarity and adjacency observed in the true comparison. Finally, the datasets are directly compared, and differential similarities are calculated. The differential similarities are segregated into two gene co-expression elasticity networks. These networks are passed through the sigmoid adjacency transform, utilizing independent parameters for the positive and negative elasticity networks. These two networks are then individually passed through the Topological Overlap process (see [121]), and hierarchical clustering is conducted to create two independent sets of gene clusters.

1. Network Comparison and Elasticity Determination

The second segment of the algorithm compares two gene expression datasets to identify statistically significant differential adjacencies on a per-edge basis. Two methods are provided for determining the cutoff for significant edge elasticity. The first method ("scalefree" extends the assumption of scale free topology in gene co-expression networks [122] to that of differential co-expression networks, and identifies parameters based on the conformity of the obtained dGCNs to the expected scale free topology criteria. The second method ("permute") utilizes random permutations of the input datasets. Fundamentally, the permutation method compares the true edgewise differential adjacency with the expected edgewise differential adjacency based on these permutations. These differential adjacencies are then translated into an elasticity matrix.

1a. Scale-free Topology Criterion

There is a great deal of evidence that shows biological networks tend to follow a scale-free distribution of node-wise edge connectivity. The first method of identifying a cutoff for significant differential adjacency is via the adherence of the produced dGCN to the scale-free topology criterion. Namely, the per-node degree distribution should adhere to a power law [122], and thereby correlate well with the Log-Log model [123].

The "scalefree" command accepts user-defined upper and lower bounds for the alpha and mu parameters. The algorithm calculates the differential similarities between all possible pairs of genes in the input dataset, and segregates these values into positive and negative elasticity networks. The algorithm then tests all combinations of alpha and mu, iterating the former by increments of 2, and the latter by increments of 0.05. For each iteration, the sigmoid adjacency function is applied, and the per-iteration distribution of per-node connectivity is then compared against the log-log model. The R-squared correlation of the per-iteration node connectivity distribution against this model is used as one of the criteria for determining appropriate parameter selection. Additional important metrics are the average per-node connectivity (the sum of all edge values associated with a given node), and the slope of the best fit linear regression. This process generates a matrix of values which show the value of the various parameters of the scale-free criterion, relative to the alpha and mu parameters.

Parameters should be chosen such that the produced positive and negative elasticity networks adhere to these criteria within acceptable limits. For the purposes of these analyses, those limits are an R-Squared correlation to the log-log degree model greater than 0.75, and a slope less than -0.8. The lowest absolute parameters (or, closest to zero) which satisfy these criteria are typically the best parameters to choose, however, the adherence of the network to the scale free criterion must be balanced against the average per-node connectivity. See [122] and http://www.danforthcenter.org/hpriest/Supplemental_File_3.xlsx

Supplemental File 4 for more information on this topic.

1b. Gene Expression Permutation

The second method of dGCN construction relies on identifying statistically significant differential similarities. In order to achieve this, the gene expression datasets are concatenated in a gene-wise fashion. The per-gene expression series are permuted via the Fisher-Yates ([107]) procedure, from which a pair of gene expression datasets (hereafter sets A and B), equal in size to the smaller of the input observation sets is extracted. For example, if two datasets A and B with 16 and 19 observations were utilized, a single set of 35 observations from a single gene would be constructed, permuted, and two randomly selected sets of observations, each of size 16 would be selected. These datasets are then passed through the dGCN construction process outlined above, to generate a dGCN based on the permuted data. The parameters utilized for calculation of the dGCN are the same as those utilized in the construction of the true dGCN. For each permutation of data, a distribution of differential similarities edge strengths is determined.

The above permutation process is repeated a user-specified number of times. This step is computationally expensive, and is made feasible by the utilization of Java's comparatively robust thread and memory management model. The resultant matrix of similarity differentials from each permutation is analyzed to build a distribution of average, per-permutation differential similarity values. This distribution, derived from random data sets, is a sampling of the expected level of random similarity variance between the specific input datasets. The true differential similarity matrix is calculated for the original input data sets, and the observed differential similarity values are compared against the average, per-permutation distribution of differential similarity values. For each differential similarity value on the interval [0,2], significance is determined by controlling for the expected false positive rate. All differential similarity values which exceed the given cutoff are tabulated in both the original datasets, and in the average per-permutation distribution. The false discovery rate is simply the average per-permutation number of edges above the cutoff, divided by the number of true edges exceeding the cutoff. The false discovery rate for every differential similarity value is tabulated and provided to the user. This process is then repeated for negative differential similarity values, on the interval [-2,0]. This process allows the user to control the number of expected false positives, versus the number of expected true positives, at all potential cutoffs. These numbers are then transformed via the sigmoid adjacency transform, and provided to the user. Once a significance level is determined for differential adjacency, this can be applied to determine elasticity networks. This is very helpful, as the two methods for identifying significant differential adjacency can be combined, to identify a cutoff that produces scale-free elasticity networks, and to assign to each produced network edge an expected false discovery rate.

2. Differential Adjacency Calculation

It is recommended to utilize both methods of estimating the correct cutoffs identifying significant elasticity. By utilizing the scale-free topology criterion to select parameters for the positive and negative elasticity adjacency transform, the user can ensure the produced networks conform to the expected topography. By utilizing the false-discovery rate estimation from permutation analysis, the user can estimate the likelihood of observing a given differential adjacency by chance – based on random permutations of the input dataset. If necessary then, the user can mask differential adjacency values they deem to be associated with insufficiently low FDR values.

Differential adjacency calculation in all applications described above and below requires the mapping of differential similarity values (lying on the interval [0,2], or [-2,0]) to the interval [-1,0] and [0,1]. This is done by dividing each differential similarity by 2. Once this mapping is complete, the adjacency function, along with the given parameters is applied. This is consistent throughout all differential adjacency calculations, so scale-free criteria and false discovery estimation all refer to common values.

3. Gene Co-Expression Elasticity Network Calculation

The resultant dGCN is further processed in two segments. The network edges representing positive elasticity (a pair of nodes which become more-co-expressed in Dataset B versus Dataset A) are treated independently of those edges representing negative elasticity (decreased co-expression in set B versus set A).

The positive and negative elasticity networks are each subjected to the topological overlap process [121]. This metric has been shown [121,124] to be a robust method of node-node association. Node pairs with many overlapping partners (i.e., they share most of their neighborhoods), have a high

TOM, whereas nodes which share no neighbors and are not connected via an adjacency edge have a TOM of zero.

The TOM-processed elasticity matricies generated by this final step of the algorithm are converted to dissimilarity matricies, and are processed through a hierarchical clustering procedure derived from [105], and two independent cluster sets, based on the pairwise dissimilarities, are generated for further analysis by the user.

Cytoscape import files are also provided. Both an unfiltered file is generated (containing the node-node differential adjacency edge strengths for all pairwise relationships) and a filtered file, containing only those edges whose differential adjacency values exceed the user-specified thresholds. In both Cytoscape files, both the differential adjacency and the absolute value of the differential adjacency edge strengths are provided.

3.2.2 GO-term enrichment analysis

All GO-term enrichment analyses described in this manuscript are carried out using the topGO R package, available via Bioconductor [125,126]. Common options in all analyses are ‘algorithm=`classic`’ and ‘statistic=`fisher`’. The GO-term to gene-locus mappings utilized for these analyses derived directly from the TAIR 10 Annotation (www.arabidopsis.org, [110]).

3.2.3 Promoter Analysis

Promoter analysis was carried out using the Element software [21]. In brief, the Element algorithm finds the rates of occurrence of all short DNA sequences (in this case, all possible sequences five to eight nucleotides in length) in the ‘background’ set of all 500 nucleotide-long upstream sequences of Arabidopsis genes in the TAIR10 annotation. Element then compares the observed occurrence rates for the same 5-8nt sequences in a set of query promoters, and identifies those sequences which are statistically over-represented. All reported statistics are corrected for multiple comparisons.

3.2.4 Individual GCN Construction

Included alongside the implementation of the dGCN analysis engine is a routine for calculation of a standard GCN. To achieve this, the software follows the popular and robust schema of the WGCNA R package [96]. Pairwise gene similarity values are calculated via a user-specified metric from the following options: the Pearson Correlation Coefficient, the GINI coefficient, and the spearman rank correlation. The sigmoid adjacency function is utilized, and a parameter determination routine ("determine") is also provided. These methods have been re-implemented in Java to facilitate full use of multithreading and a streamlined memory. In the case of the Columbia-0 network and the Lhy-OX network, parameters a=20 and mu=1 were utilized. In the unified data analysis, alpha = 30 and mu = 0.9 was utilized.

Parameter Determination

It is critical to determine appropriate parameters for the sigmoid adjacency function, and the implementation of the algorithm includes a routine specifically targeted at achieving this.

Ideally, the slope of the linear regression should be near -1, and the R-squared correlation should be greater than 0.75. It is best to select parameters that satisfy these criteria without sacrificing high average node connectivity. More information and guidance on this process is available in

http://www.danforthcenter.org/hpriest/Supplemental_File_3.xlsx

Supplemental File 4.

3.3 Results

In order to demonstrate that dGCNA identifies biological signals that cannot be identified through a classic GCN approach, we analyzed two circadian time-courses in the model organism *Arabidopsis thaliana*. Full growth conditions, sampling protocols, RNA preparation, array hybridization, array quality control, and array normalization procedures are described in [127]. In short, Wild-type Columbia-0 (Col-0), and a line over-expressing the core circadian clock oscillator LATE ELONGATED HYPOCOTYL (LHY, AT1G01060) (Lhy-OX), were subjected to circadian entrainment via short-day (SD) photo/thermo-cycles for 7 days. Samples were taken every four hours for 48 hours, beginning at subjective dawn. Arrays were RMA-normalized utilizing the Affymetrix software, and per-sample normalized gene expression values were produced.

Each time-course was described by a set of 25,000 gene expression data series, of 12 time-points for each genotype. In order to be included in the set of probe-sets to be analyzed, a probe-set must have at least one data point with non-log RMA normalized intensity greater than 50, in at least one of the two data sets. This generated a list of 9,882 probe-sets. Although the algorithm and implementation described herein easily handles gene sets of much larger size, in practice it is not always ideal to use the largest possible gene input sets. Low-variance genes and low expression genes may only expand computational resource requirements and complicate secondary analyses, while adding little or no information of use.

In all subsequent analysis, these datasets were analyzed in three ways. First, as independent datasets, the Columbia-0 time course (“Col-0”), the *At*LHY over-expression time-course (“Lhy-OX”). Secondly, as a ‘unified’ dataset, in which the data series from the Col-0 and Lhy-OX datasets were concatenated in a gene-specific fashion (“Unified”). Finally, as a comparative dataset, in which the algorithm described herein compared the datasets against one another. This final analysis generated a total elasticity network (all statistically significant elastic gene-gene edges), and two subnetworks: the negative elasticity network (all gene-gene relationships which experience a decrease in co-expression from Col-0 to Lhy-OX), and the positive elasticity network (all gene-gene relationships that experience an increase in co-expression from Col-0 to Lhy-OX). The Pearson Correlation Coefficient (PCC) was utilized for all analyses of these data.

3.3.1 Differential Gene Co-expression Network Analysis

The total elasticity network was calculated by first utilizing the 'scalefree' command. All combinations of values for the mu parameter on the interval [0.5,1] and the alpha parameter on the interval [16,30] were tested. Results are shown inhttp*://www.*danforthcenter.org/hpriest/Supplemental_File_4.xlsx

Supplemental File 5. Based on these results, parameters of alpha = 26 and mu = 0.8 were selected for the positive elasticity network, and parameters of alpha = 28 and mu = 0.8 were selected for the negative elasticity network.

Adjacency cutoffs for dGCN construction were determined estimation of statistical significance of each produced differential adjacency value. The 'permute' command was run with the same sigmoid adjacency function parameters identified above, with 100 permutations. The distribution of per-permutation observed differential similarity values was computed, and the observed true differential similarity values are compared against this background. These differential similarity values are then transformed via the adjacency transform. The cutoffs of positive differential adjacency >= 0.01, and negative differential adjacency <= -0.01 correspond to estimated false-discovery rates of FDR < 0.0252 and FDR < 0.0380, respectively. These produced cutoffs apply to network adjacencies – not differential similarities. In this particular case, both methods utilized for determining the final network structure gave rise to highly similar networks. A new set of permutations produced essentially the same differential similarity cutoffs, within 0.02 similarity. Larger permutation sets (200, 300, 500) did not reduce this variance.

By combining both the determined best sigmoid adjacency parameters for positive and negative elasticity calculation, as well as the cutoffs for statistical significance, the overall elasticity network could be calculated. With the 'compare' command, the two datasets were directly compared. The overall differential network was then masked to isolate the two sub-networks, the positive (edge-specific differential-adjacency >= 0.01) and negative (edge-specific differential-adjacency <= -0.01) elasticity networks. Individual networks calculated from each dataset (see methods), comprised 18,739,031 total unique edges. In total, the elasticity networks comprise 4,368,745 edges. This represents 23.31% of edges which were found to exist which, and 4.47% of the 97,653,924 total possible edges.

The positive elasticity network (scale free criterion: 0.774, slope: -1.117, mean connectivity: 99.45) contains 2,932,647 edges (67.1% of all elastic edges). The negative elasticity network (scale free criterion: 0.755, slope: -1.115, mean connectivity: 91.02) contains the remaining 1,436,098 edges (32.9% of total elastic edges). The positive and negative elasticity networks, made up only of significantly differential adjacency edges, were subjected to hierarchical clustering and produced cluster sets of 65 and 61 clusters, respectively.

3.3.2 Comparison of Elasticity Networks to Standard GCNs

It is critical to establish that the algorithm presented here generates novel insight into the gene-gene relationships of a particular set of data. We must determine that the elasticity analysis presented identifies biological meaning distinct from that revealed by previous approaches. It is important to consider the source of the data – a pair of circadian time-courses, from wild-type

Columbia-0 and a line over-expressing the core circadian oscillator AtLHY (AT1G01060, [128]). In the comparisons presented, we would expect that the circadian clock would be mis-expressed, and that effects of AtLHY overexpression would be evident in aspects of gene expression relating to abiotic stress response (which is regulated by the circadian clock) [129], energy harvesting (which is phased to specific times of day to optimize growth and energy harvesting)[130], light-sensing (in which AtLHY is a principle factor)[128], and growth itself (for which AtLHY has been shown to be a critical regulator)[131].

Network Level Analysis

It is important to distinguish the produced elasticity networks from standard GCNs. We wished to determine if the elasticity networks describe a substantially different set of relationships from that of classic GCNA. We therefore directly compared the sets of gene-gene relationships which were identified utilizing a classic GCNA approach, and those edges identified as having statistically

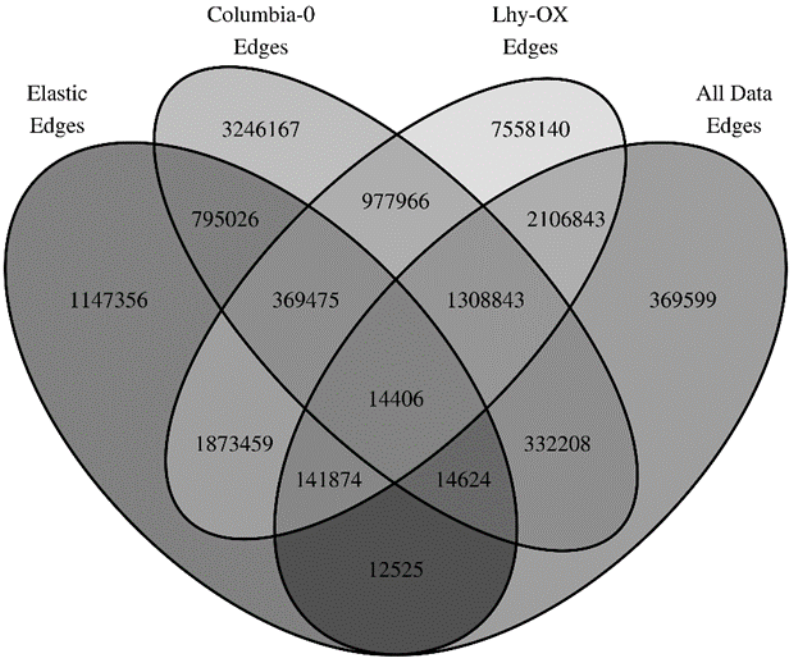


Figure 7 Overlaps between gene-gene edge sets. Overlaps were identified between the total elasticity, unified, *At*LHY-OX, and Columbia-0 networks. Edge sets were determined by binary presence/absence of an edge between gene pairs.

significant elasticity. If dGCNA reveals novel biological signals, both the set of edges which are represented in the network, as well as their edge values, should be highly distinct from the edge sets of a classic GCNA.

Again, standard GCNs were calculated for three datasets: the original two datasets, consisting of data from Col-0 and Lhy-OX, and the union of the two datasets (the Unified dataset), in which each gene is represented by a total of twenty-four observations. These sets represent, in our view, the most common approaches to network analysis of multiple datasets such as these. The overlap of the edge sets contained within these three networks and the total set of statistically significant elastic edges is shown in **Figure 7**.

The individual GCNs constructed from the Col-0 and Lhy-OX datasets contained 7,058,715 and 14,351,006 edges, respectively. Of those, 2,670,690 edges were shared between them. Of these,

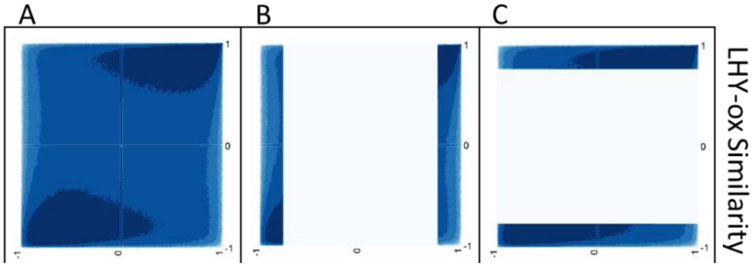


Figure 8. Edges identified by dGCNA are distinct those identified using previous approaches. (A-E) Density heatmaps of edge values for different sets of edges. The x-coordinate of an edge is determined by that edge's value in the Columbia-0 network. The y-coordinate of an edge is determined by that edge's value in the *At*LHY-OX network. Edges falling along the x=y line have similar values in both datasets. (A) Density heatmap for all possible pairwise edges, totally 97,574,884 pairings, demonstrating the possible 'edge-space' of the two datasets. (B) Density heatmap for edges in the Columbia-0 network. (C) Density heatmap for edges in the *At*LHY-OX network. (D) Density heatmap for edges in the Unified data network. Majority of edges fall near the x=y=1 or x=y=-1 region, indicating the Unified analysis approach captures relationships that are stable between both datasets. (E) Density heatmap for edges in the total elasticity network. dGCNA identifies those edges which have very strong changes in value between the underlying datasets. All edges included in (E) have FDR <= 0.0380, see methods.

only 383,881 edges (14.4%) were also shared with the total elasticity network. Elasticity analysis identified 795,026 edges previously unique to the Col-0 network as elastic (19.7% of Col-0-specific edges), and 1,873,459 of 9,431,599 of edges previously unique to the Lhy-OX network (19.9% of Lhy-OX-specific edges). In order to better assess both the set of edges which are included in each network, and the strength of those edges, each network's edges had their values plotted as a heatmap.

Figure 8A shows the distribution of all possible 97,574,884 edges based on their values in the Col-0 and Lhy-OX datasets. Most edge values tend to cluster in the upper right or lower left quadrants, with a great deal of spread, as is expected. The individual networks calculated from the total, unfiltered similarity network (**Figure 8, B and C**) show the practical effect of adjacency transforms in setting a lower limit for acceptable similarity. In **Figure 8B**, the Col-0 edge set shows a uniformly strong set of relationships (all edge values near -1 or +1), whereas there is no discernable selection for values in the Lhy-OX dataset. In **Figure 8C**, the Lhy-OX edge set shows strong relationships, with all edge values near -1 or +1, and there being no discernable dependence on edge value in Col-0. **Figure 8D** shows the edges included in the unified data set network. Many edges have entirely un-remarkable edge values in Col-0 or Lhy-OX, falling near the origin of the plot. However, there is a very strong bias for edge values which are close to +1,+1, and -1,-1. This tendency reveals the pitfall of performing GCNA on combined datasets. The majority of edges included in the Unified network have strong relationships in both of the included subnetworks. A great many edges that were included in the subnetworks are not included in the unified dataset. Finally, **Figure 8E** shows the values of the edges included in the elasticity network. These edges have strongly differing edge values in the original Col-0 and Lhy-OX datasets. This edge set is highly distinct from any of those shown in **Figure 8B-D**. All of the edges identified to have statistically significant elasticity change their edge value drastically between the Col-0 and Lhy-OX networks. Although some of the elastic edges are identified in either of the two original data-sets, the gene relationships these edges represent are identified as undergoing significant changes in response to the genetic perturbation. Of the 4,368,745 elastic edges, 1,147,356 (26.3%) were entirely novel, which indicates that without comparing the two original datasets, there would be no reason to expect the genes represented by those edges were related at all. Of the remaining elastic edges, 2,668,485 were shared with either the Col-0 or Lhy-OX networks only, identified portions of those networks as of particular interest in their response to the genetic perturbation. Only 183,429 elastic edges were shared with the unified network (4.2%). This particularly low overlap shows the utility of viewing a pair of datasets in such contrast. It is clear that dGCNA reveals novel gene-gene relationships, as well as shows the elasticity of previously identified gene-gene relationships in response to genetic perturbation.

Network Analysis by Node Connectivity

As our differential adjacency analysis identifies node-pairs which undergo significant changes in their relationship, we hypothesized that genes proximal to the genetic perturbation in the regulatory landscape of *Arabidopsis* would be more likely to undergo elasticity. We would expect that those nodes with many significant elastic connections would be associated the functions for which *At*LHY is known to regulate or be associated with. We therefore identified two groups of nodes: the 5% of nodes with the highest total connectivity (K), and the 5% of nodes with the lowest total K. These connectivity groups were identified for each of the networks generated: the total elasticity network, the positive and negative elasticity networks, the individual Col-0 and Lhy-OX networks, and the Unified data network. Because each network contained approximately 9800 genes, these groups numbered 490 genes each. Each group was subjected to GO-term enrichment analysis. **Table 6** contains the GO-terms which were statistically over-represented in the group of high-K nodes from the total elasticity network. Any GO-terms which were also found to be over-represented in the high-K groups in any of the three individual networks (Col-0, Lhy-OX, or Unified) are not listed. The GO-terms were found to only be statistically over-represented in the elasticity analysis correspond very closely with those functions *At*LHY is known to regulate. The GO-terms represent abiotic stress: GO: GO:0009266, “response to temperature stimulus”, GO:0009409, “response to cold”, GO:0006970, “response to osmotic stress”, and GO:0009651, “response to salt stress”. The enriched terms also represent light-sensing and growth: GO:0009416, “response to light stimulus”, GO:0009314, “response to radiation”, GO:0010051, “xylem and phloem pattern formation”, and GO:0007389,“pattern specification process". The complete lists of statistically enriched GO Terms for each connectivity group are provided in ***http://www***.danforthcenter.org/hpriest/Supplemental_File_5.xlsx

Table 6. Gene Ontology Terms Found to be Statistically Over-represented Only in the Elasticity Networks produced by dGCNA.

Gene Ontology Term	Short Description	Total Elasticity	Neg. Elasticity	Pos. Elasticity	Col-0	Lhy-OX	Unified
GO:0009266	response to temperature stimulus	9.85E-08	0.0017	0.0003	0.8557	1.0000	1.0000
GO:0009409	response to cold	9.96E-07	0.0022	0.0011	1.0000	1.0000	1.0000
GO:0007623	circadian rhythm	1.90E-06	4.70E-06	3.01E-05	0.5784	1.0000	1.0000
GO:0048511	rhythmic process	1.90E-06	4.70E-06	3.01E-05	0.5784	1.0000	1.0000
GO:0006970	response to osmotic stress	6.22E-05	0.0030	0.0655	0.8201	0.5419	0.9769
GO:0009651	response to salt stress	0.0005	0.0073	0.2124	0.6945	0.4238	0.8207
GO:0009416	response to light stimulus	0.0007	0.0870	0.0043	1.0000	0.0507	1.0000
GO:0009314	response to radiation	0.0011	0.1123	0.0055	1.0000	0.0626	1.0000
GO:0005983	starch catabolic process	0.0011	0.0017	0.0018	1.0000	1.0000	1.0000
GO:0015994	chlorophyll metabolic process	0.0013	0.0019	0.0004	0.8201	0.4493	1.0000
GO:0044247	cellular polysaccharide catabolic process	0.0014	0.0019	0.0021	1.0000	1.0000	1.0000
GO:0009737	response to abscisic acid stimulus	0.0017	0.1449	0.0277	1.0000	1.0000	1.0000
GO:0009251	glucan catabolic process	0.0017	0.0022	0.0027	1.0000	1.0000	1.0000
GO:0009415	response to water	0.0019	0.0075	0.0060	0.8022	1.0000	1.0000
GO:0006778	porphyrin metabolic process	0.0025	0.0008	0.0034	0.4332	0.7119	1.0000
GO:0044275	cellular carbohydrate catabolic process	0.0026	0.0008	0.0034	1.0000	1.0000	1.0000
GO:0033013	tetrapyrrole metabolic process	0.0026	0.0008	0.0034	0.4332	0.7119	1.0000
GO:0009725	response to hormone stimulus	0.0037	0.2167	0.0980	0.4332	0.8434	0.0519
GO:0009719	response to endogenous stimulus	0.0037	0.2251	0.1024	0.4332	0.8545	0.0536
GO:0006807	nitrogen compound metabolic process	0.0040	0.0214	0.0020	1.0000	1.0000	0.8937
GO:0044283	small molecule biosynthetic process	0.0041	0.4793	0.0514	0.8492	0.0622	1.0000
GO:0015995	chlorophyll biosynthetic process	0.0070	0.0097	0.0018	1.0000	0.5419	1.0000
GO:0009631	cold acclimation	0.0074	0.0803	0.0655	0.3914	1.0000	1.0000
GO:0006779	porphyrin biosynthetic process	0.0098	0.0023	0.0119	0.5008	0.8434	1.0000
GO:0034641	cellular nitrogen compound metabolic process	0.0101	0.0524	0.0055	1.0000	1.0000	1.0000
GO:0009414	response to water deprivation	0.0104	0.0460	0.0360	0.7544	1.0000	1.0000
GO:0000272	polysaccharide catabolic process	0.0104	0.0166	0.0138	1.0000	1.0000	1.0000
GO:0005982	starch metabolic process	0.0108	0.0007	0.0660	1.0000	1.0000	1.0000
GO:0033014	tetrapyrrole biosynthetic process	0.0108	0.0030	0.0146	0.5396	0.8906	1.0000
GO:0010051	xylem and phloem pattern formation	0.0207	0.4765	0.0243	1.0000	1.0000	1.0000
GO:0016070	RNA metabolic process	0.0212	0.4793	0.0116	1.0000	1.0000	1.0000
GO:0007389	pattern specification process	0.0216	0.0910	0.1818	1.0000	1.0000	1.0000
GO:0003002	regionalization	0.0220	0.2910	0.0730	1.0000	1.0000	1.0000
GO:0044271	cellular nitrogen compound biosynthetic process	0.0247	0.0022	0.0055	0.4332	0.4287	1.0000
GO:0016052	carbohydrate catabolic process	0.0305	0.0034	0.4775	1.0000	0.2832	1.0000
GO:0009845	seed germination	0.0438	0.0612	0.3701	1.0000	1.0000	1.0000
GO:0009058	biosynthetic process	0.0482	0.3856	0.0519	1.0000	0.2651	1.0000
GO:0042440	pigment metabolic process	0.0482	0.0182	0.0153	0.1599	0.5647	1.0000
GO:0046148	pigment biosynthetic process	0.0495	0.0182	0.0153	0.1828	0.6459	1.0000

Supplemental File 6.

In addition to the above analysis, we analyzed the promoters of each of the above sets for over-representation of DNA motifs. The promoters of the high-K grouping derived from the total elasticity network were statistically over-represented for the DNA elements: "AAATATC", “AAATATCT”, “AATATC”, “ATATC”, “AATATCT”, and “AAAATATC”. All elements were in the top 10 hits (maximum FDR-corrected p-value < 3.34e-09). These elements are all near-exact matches for the Evening Element, which *At*LHY is known to bind. The total sets of all over-represented DNA motifs, for all high- and low-K groups are contained in http://www.danforthcenter.org/hpriest/Supplemental_File_6.xlsx

Supplemental File 7.

3.3.3 Biological Meaning of Elasticity Networks

Cluster-level Analysis

At the highest level, we would expect to see each cluster of genes identified as positively and negatively plastic undergo changes in co-expression between networks. The dataset-specific expression profiles of the genes of each module were plotted for direct comparison.

The genes of each cluster were subjected to GO-term enrichment analysis. It would be expected that the large changes in network topology triggered by over-expression of AtLHY would be related to the general functions which AtLHY regulates. We see that this is the case.

Every gene can be a member of one module in the positive elasticity network, and one module in the negative elasticity network. We highlight here two modules, the first, module 8 of the negative elasticity network (module -08), and the other, module 1 of the positive elasticity network (module +01). These modules are highlighted for their combination of biologically interesting GO-term enrichment and expression profiles.

Module -08 has 321 member genes, which are enriched for several abiotic stress response GO terms (GO:0009409, GO:0009414, “response to cold”, “response to water deprivation”, respectively, both FDR-corrected p-value <0.02), as well as GO:0015979, “photosynthesis”, FDR-corrected p-value < 0.0035. Genes of this module tend to have a very strong spike in expression at zt08 in the Columbia-0 wild-type (**Figure 9A)**. Under the growth conditions of the experiment, this corresponds with lights-off. Under the Lhy-OX perturbation, however, this coordinated increase in expression is phased to zt00/zt24, and is much attenuated (**Figure 9B**). Many of the genes never spike in expression at all. Element analysis of the promoters of the gene members of module -08, revealed that the exact evening element core (TATC) appears in six of the top twenty most enriched sequence elements (all FDR-corrected p-value < 3.5 x10^{-6}). These results are obtained with no *a priori* assumptions or inputs regarding the function of AtLHY. An additional eight of the top twenty elements contain partial matches to the evening element.

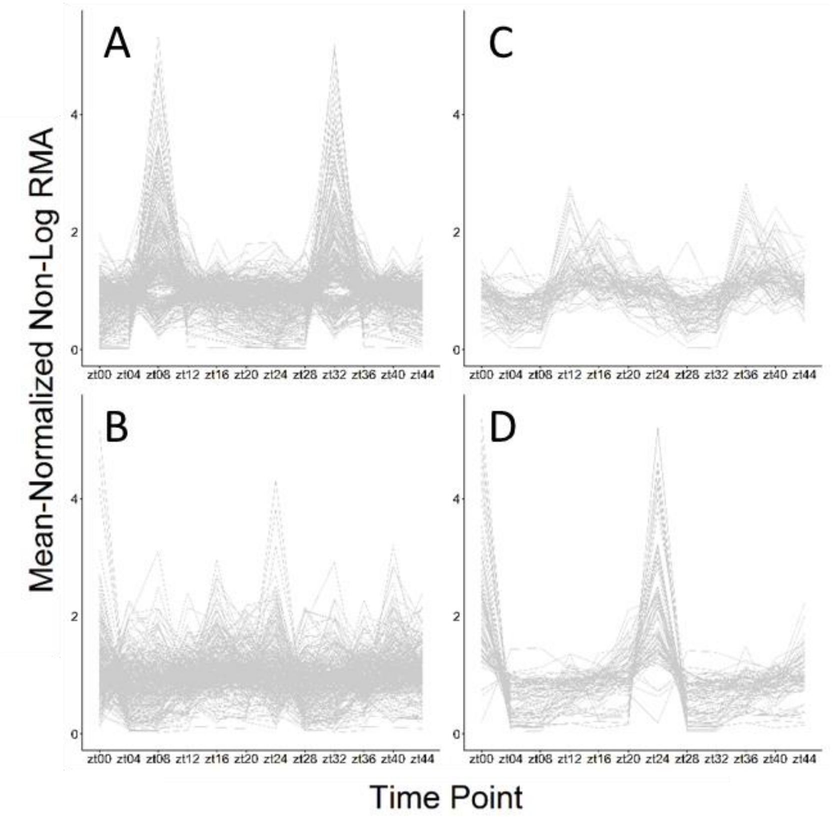


Figure 9. Mean-normalized, RMA-normalized gene expression for two modules of genes undergoing significant elasticity. The RMA-normalized expression values for each gene were normalized by the mean expression value for that gene. Gene expression values (y-axis) for the genes in a particular module were plotted to give an indication of that modules overall expression pattern over the course of a short-day circadian experiment (x-axis). (A) Expression pattern for genes of module -08 based on data derived from the Col-0 dataset. (B) Expression pattern for the same set of genes depicted in (A), based on expression data derived from the *At*LHY-OX dataset. Module -08 undergoes significant negative elasticity, exemplified here by the loss of the distinct expression peaks at zt08 and zt32. The genes are loosely co-expressed under over-expression of *At*LHY, with almost no apparent coordinated expression pattern. (C) Expression pattern for genes of module +01 based on data derived from the Col-0 dataset. (D) Expression pattern for the same set of genes depicted in (C), based on expression data derived from the *At*LHY-OX dataset. Module +01 undergoes significant positive elasticity, in which a set of genes that is loosely co-expressed in a wild-type background becomes very strongly co-expressed under over-expression of *At*LHY.

Module +01 has 81 member genes, which are strongly enriched for GO-terms related to stress and abiotic stress, chemical and carbohydrate stimulus, as well as several GO-terms relating to biotic stress responses. These systems are all known to be regulated to various degrees by the circadian system [132]. Module +01 is also of great interest for the dramatic change in expression profiles of the constituent genes. Genes of module +01 experience moderate co-expression, with general day-time repression (zt00 through zt08) followed by night-time expression at multiple time-points (zt12 through zt24, **Figure 9C**). Almost all gene members of module +01 experience a sharp expression peak at dawn (zt00 and zt24) under overexpression of *At*LHY (**Figure 9D)**. The promoters of these member genes are not substantially enriched for any particular known circadian or stress related element, though there are some DNA motifs which are enriched that are weak matches to part of the Evening Element. There is no clear hypothesis to draw regarding the transcriptional regulation of the dramatic shift in expression profile and co-expression of the genes of module +01.

The full module gene lists, the complete listing of all promoter analysis results, and each module's GO-term enrichment statistics may be found in

http://www.danforthcenter.org/hpriest/Supplemental_File_7.xlsx

Supplemental File 8.

Table 7. Enrichment of the Evening Element in the Immediate Neighborhood of AtCCA1 and AtLHY.

		Evening Element	
Neighborhood	# Promoters	Motif Variants	Highest Rank
LHY Total Elasticity	625	21	12
LHY +Elasticity	401	18	4
LHY -Elasticity	224	1	31
LHY Col-0	1293	8	51
LHY LHY-OX	2354	7	226
LHY Unified	956	6	36
CCA1 Total Elasticity	2069	9	74
CCA1 +Elasticity	1096	1	102
CCA1 -Elasticity	973	3	90
CCA1 Col-0	1304	4	94
CCA1 LHY-OX	1968	5	169
CCA1 Unified	122	1	26

Gene Level Analysis

In order to determine if the individual gene-gene relationships which comprise the elasticity networks hold meaning, we examined *At*LHY's immediate differential adjacency network. *At*LHY had 625 differential edges, 401 positive, 224 negative. The Evening Element (EE) core sequence (TATC/GATA) is a known binding site of *At*LHY. We surmised that if the genes in the immediate neighbor of *At*LHY are in fact under the regulatory control of *At*LHY to some degree, they should be enriched for the EE core sequence. **Table 7** contains the results of Element analysis of the promoters of the genes in the immediate neighborhoods of *At*CCA1 and *At*LHY. Analysis of the promoters of genes immediately proximal to *At*LHY in the overall elasticity network revealed 21 DNA motifs containing the EE core. We wished to determine if these elements were selectively enriched in either the positive or negative elasticity networks. Enrichment of the evening element was distinctly segregated. 18 EE core containing DNA elements were over-represented in the positive elasticity network, and only one such DNA element was enriched in the negative elasticity network. The element ATATC was found 828 times, in 325 of 401 promoters of positive elasticity genes connected to *At*LHY (FDR-corrected p-value < 1.72x10-^{17}). Motifs containing exact matches of the EE make up 6 of the top 10 most enriched motifs. Similar analysis on the immediate neighborhood of *At*LHY in the individual networks do not reveal similar signal. The EE core sequence first appears as the 226th most-enriched in the promoters of *At*LHY's immediate neighbors in the LHY-OX network. The next closest appearance is 252nd. Discovery of *At*LHY's binding target fares little better in the ColSD network, with the highest appearance occurring at 51st overall. Variants of the EE core appear only 7 times in DNA elements over-enriched in the immediate neighborhood of *At*LHY in the Lhy-OX network, and only 8 times total in the promoters of *At*LHY's neighborhood in the ColSD network. We next investigated if the enrichment is related specifically to *At*LHY connectivity, or simply enriched in genes associated with *At*LHY through its regulation of the circadian system. We identified the immediate neighbors of *At*CCA1. *At*CCA1 forms a heterodimeric transcription factor complex with *At*LHY (citation needed). Only 9 total variants of the EE core are found in the total elasticity network of *At*CCA1, and the highest-ranked sequence appears at 74th overall. Elasticity analysis does not appear to enrich for this signal in the neighbors of *At*CCA1, with the presence of the EE core sequence being roughly the same in the elasticity networks as it is in the individual networks. In-*vivo* work would need to be completed to ascertain which of the putative 325 promoters identified by the elasticity analysis *At*LHY actually binds to, but it would appear that network comparison and elasticity analysis isolates biological signal quite well.

The immediate neighbors of *At*LHY are also significantly enriched for GO terms of interest, including GO:0007623, "circadian rhythm" in the top 5 terms of the negative elasticity network, and GO:0009409, "response to cold" in the top 5 terms of the positive elasticity network. All enriched terms in immediate neighborhood of *At*LHY in the elasticity networks are included in ***http://www***.danforthcenter.org/hpriest/Supplemental_File_8.zip

Supplemental File 9. The above terms are also enriched in the immediate neighborhoods of *At*LHY in the individual networks. However, in the elasticity network, these terms are contained in a list of 83 total terms, the majority of which represent functions which are regulated by the circadian system. In the individual networks, the number of enriched terms is 483 – dGCNA in this case clearly isolates a much stronger biological signal.

3.3.4 Comparison to DiffCoEx

DiffCoEx [133] is an algorithm primarily concerned with identifying differential co-expression on a modular level. This enables the analysis of how large groups of genes behave as groups, but does not enable in any way the analysis of genes on an individual level. In other words, DiffCoEx determines if the genes of module A change in co-expression with the genes in module B, and conducts that comparison for all possible pairings of modules. dGCNA identifies statistically significant co-expression changes on a gene-to-gene level, and builds module sets that reflect those changes.

DiffCoEx identified 24 modules of genes – however these groups are not broken out into groups that increase in co-expression and groups that decrease in co-expression, but are an agglomeration of both behaviors. The module set produced by DiffCoEx encompassed all genes in the analysis. It is worth noting, especially in light of the analysis below, that DiffCoEx and dGCNA are fundamentally different in their approach to differential co-expression in high-dimensional gene expression datasets.

In order to compare the modular analyses of DiffCoEx and dGCNA, we compared the gene member lists of all three independent modular sets against each other; the DiffCoEx set, the dGCNA 'positive elasticity' set, and the dGCNA 'negative elasticity' set. If the modular analyses gave largely the same groupings, a gene list from one set of modules should have strong overlap with a gene list from another set of modules. This is easily determined by an overlap-based enrichment analysis. This occurred infrequently in the comparisons we conducted with only one DiffCoEx module overlapping with a single dGCNA module (royalblue3 with module -43, p < 0.001). In fact, the average number of significant overlaps between a DiffCoEx module and a dGCNA module was 7.52. The 'mediumorchid' module overlapped significantly with 36 dGCNA modules – more than a quarter of all modules identified by dGCNA. Similar analysis found that dGCNA modules tended to significantly overlap with multiple DiffCoExp modules – this would imply that it is not simply the case that dGCNA finds submodules of DiffCoEx modules, or vice-versa. Overall, the modular analyses overlap only weakly. This reflects the major differences in the underlying analysis targets, and approaches, of DiffCoEx and dGCNA.

The modules identified by DiffCoEx were subjected to the same Element-based promoter analysis as those modules identified by dGCNA. The modules were also subjected to GO-term enrichment analysis. *At*LHY was a member of the 'darkolivegreen2' module, a module with 139 member genes. These genes were over-represented for three GO-terms of great interest: GO:0007623, 'circadian rhythm', GO:0048511, 'rhythmic process', GO:0042754, 'negative regulation of circadian rhythm' (FDR-corrected p-value < 0.25, 0.25, and 0.27, respectively). In addition to three other GO-terms: GO:0042221, 'response to chemical stimulus', GO:0046685, 'response to arsenic', and GO:0051179, 'localization' (all FDR-corrected p-value <0.027). The promoters of this module were enriched for twenty DNA motifs, however, most were weak matches to a highlyredundant portion of the evening element (“AAATAA”), and only two were partial matches to the G-box (“CACGTA”, and “ACGTG”, both FDR-corrected p-value < 2.4x10^{-4}).

In the elasticity analysis done via dGCNA, AtLHY is a member of module 28 in the positive elasticity network (+28) and module 30 in the negative elasticity network (-30). Module +28 is not statistically enriched for any GO terms. However, module -30 is statistically enriched for three GO terms: GO:0009416, “response to light stimulus”, GO:0009628 “response to abiotic stimulus”, and GO:0009314, “response to radiation”, all at FDR-corrected P-value < 0.005. These three gene ontology terms represent primary functions which AtLHY has been shown to regulate, and are the only terms for which module -30 is enriched. The promoters of this module were enriched for 48 DNA elements, including many partial matches to the G-Box sequence, as well as an exact match (“CACGTG”, FDR-corrected p-value 8.9x10^{-5}).

DiffCoEx does not provide access to node-node interactions, or a tabulation of each node-node interaction that is statistically significantly different. This is a major advantage provided by dGCNA, and our analysis above indicates that these node-node interactions, specifically when they denote elasticity, may contain important biological signals which are excluded by DiffCoEx. While biological experimentation would be required to determine which module set captures more accurate biological signal, there is no question that dGCNA provides a more granular analysis of node-node interactions in the differential co-expression relationships of genes.

3.4 Discussion

The comparison of gene co-expression networks is an underutilized area of gene expression analysis. With sequencing prices continuing to drop, and yields increasing, it is easier than it has ever been to conduct a complex, multi-dimensional gene expression experiment. Computational biology method development must keep pace with the commonality of increased complexity in gene expression datasets. With dGCNA, presented here, we show that the comparison of GCNs reveals novel biological signals on three levels: the whole-network level, the gene-cluster level, and the individual gene-to-gene relationship level.

3.4.1 Novel Whole-Network Biological Insights Revealed by Elasticity Analysis

Network-level gene co-expression trends have not been previously shown to hold strong biological signals. While network-wide connectivity distributions are an important part of network reconstruction [96,122], and module-level connectivity plays an important role in prediction of criticality of individual genes, the connectivity of a gene across an entire GCN has not previously been of great interest. In our analysis the grouping of genes which are most elastic – that is, they possess the greatest number of statistically significant elastic edges – are directly related to the functions regulated by *At*LHY (**Table 6**). In addition, the promoters of those highly elastic genes are very strongly statistically enriched for the Evening Element, which is the known binding site of *At*LHY (***http://www***.danforthcenter.org/hpriest/Supplemental_File_6.xlsx

Supplemental File 7).

This is analogous to a direct differential gene expression analysis. By employing a comparative strategy, dGCNA identifies only those connections which change significantly between conditions. A high-connectivity gene node, which is un-related to the perturbation under study would be removed from an elasticity analysis. While employing GCNA allows the discovery of broad trends of co-expression within a dataset, applying dGCNA with a pair of datasets allows discovery of only the differences of trends between datasets, and avoids capture of what could be termed ‘house-keeping’ trends.

3.4.2 Cluster Level Analysis

The generation of gene clusters which are similarly expressed across a large dataset has always been a key strength of GCNA. Elasticity analysis continues this, by enabling the tracking of large-scale changes in gene co-expression between datasets. **Figure 9** exemplifies this facet of dGCNA. Under overexpression of *At*LHY, a number of previously phased to the day/night transition are mis-expressed. These genes are strongly enriched for photosynthesis and stress-response related genes (GO:0009409, “response to cold”, FDR-corrected p-value <0.02 and GO:0015979, “photosynthesis”, FDR-corrected p-value < 0.0035). Light/Dark transitions are the principle method of entrainment for the diurnal/circadian system, and *At*LHY is a principle mediator of that pathway. Many genes are tightly regulated to be expressed a specific time of day, to best adapt the organism to the surrounding environment. This particular module of genes is expressed at dusk (**Figure 9A**), and, under *At*LHY over-expression, lose their co-expression almost entirely. Their spike in expression at zt08 is entirely lost, and the genes which still undergo large changes in expression do so at zt00 – dawn – instead of dusk (**Figure 9B**).

Module +01, on the other hand, has weak-to-moderate co-expression, being principally expressed during the dark period (**Figure 9C**) in the Columbia-0 control. Under over-expression of *At*LHY, the genes of module +01 experience a drastic change in expression profile, becoming highly expressed at dawn (**Figure 9D**). Our algorithm is very effective at identifying not only gene-to-gene co-expression changes, but identifying groups of genes which all change co-expression in similar fashions. In this way, the commonalities of gene-gene relationships between a pair of datasets can be removed, and stark, broad changes such as those shown in module +01 are revealed.

3.4.3 Meaning of node-node interactions in elasticity network

To date, there is little information about the true meaning of gene-gene interactions in GCNs. The edges in GCNs are often mistaken to imply a molecular interaction (i.e., gene A regulates gene B), or a similar physical connection of proteins. However, edges in GCNs denote co-expression – nothing more. The meaning of such a relationship is contextually rooted in the underlying data. The forgoing notwithstanding, our analysis of the local connections of *At*LHY in the elasticity network suggest more meaning in a dGCN. In the total elasticity network, the promoters of the immediate neighbors of *At*LHY were substantially enriched for the Evening Element – again, the direct binding substrate of *At*LHY (**Table 7**, Row "LHY Total Elasticity"). In comparison to standard GCNA of the individual and unified datasets, the over-representation of the Evening Element in neighbors of *At*LHY is very strong, with the evening element occurring less than half as many times, in twice as many promoters (**Table 7**, Rows "LHY Col-0", "LHY LHY-OX", and "LHY Unified"). The substantial enrichment of the EE is found only by application of dGCNA. Further molecular work would need to be carried out to confirm if *At*LHY does in fact bind to any of the promoters of its gene neighbors. However, from a strictly *in-silico* standpoint, the results are excellent.

3.4.4 Comparison to Similar Methods

The complexity of both networks and network comparison has generated a number of software solutions [116–119]. These solutions, while effective, do not apply to the identification of statistically significant differential adjacencies within large, transcriptomic-scale gene expression datasets.

The underlying data sources, targeted products, and scale of dGCNA are closely aligned with the DiffCoEx procedure [120]. The comparison to DiffCoEx revealed fascinating results. Each algorithm was able to associate with *At*LHY several of the functions that *At*LHY is known to regulate or influence. However, each also failed to capture the entirety of the functional space of *At*LHY. In fact, the results are non-overlapping, with DiffCoEx identifying the rhythmic and circadian associations of the transcription factor in question, and dGCNA identifying light-sensing and abiotic stress responses as the principle related functions. Neither algorithm was able to overwhelmingly associate putative transcription factor binding sites with the genes that co-express, or differentially co-express with *At*LHY between the Col-0 and Lhy-OX datasets. In addition, the modular analyses of dGCNA and DiffCoEx are highly non-overlapping. This is certainly a direct result of the entirely independent approaches towards analysis of the changes in co-expression between datasets. DiffCoEx targets identification of module-to-module differential expression. The target of dGCNA is to assess significant gene-to-gene changes in differential co-expression, and to build modules from those changes.

This allows dGCNA to provide both a broader, and a more granular view of differential co-expression than DiffCoEx. Individual gene-gene differential co-expression relationships are provided directly to the user. As we have demonstrated above, these relationships carry strong biological function, indicating in this case both putative regulatory circuitry and functional relatedness. In addition, dGCNA carries out the process of its algorithm in an automated fashion, without requiring the user to manually manipulate the data at each algorithm step. Simultaneously, the user retains the ability to fully customize each step of the algorithm, enabling great ease-of-use without sacrificing power or customization.

3.5 Conclusions

Gene co-expression elasticity is an expected property of biological systems. Transcriptional regulation is multi-layer system consisting of pre-, co-, and post-transcriptional regulatory effects, all of which change in response to external stimuli. Analytical methods that treat co-expression relationships as concrete objects have been shown to be severely limited when applied to large-scale, multi-experiment expression datasets [134]. These approaches begin to miss biological signal when applied to even relatively small scale, but independently generated expression datasets [69].

Here we have presented dGCNA, implemented in java, which enables investigators to make inquiries into the co-expression landscape of complex, large-scale expression datasets. Our algorithm enables the tracking of individual co-expression relationships between conditions, thereby allowing observation of GCN remodeling in response to stress, circadian entrainment, genotypic diversity, developmental stage, or any perturbation of interest. We show that our method is a significant improvement over previous computational methods in the area, allowing far finer understanding of co-expression elasticity, as well as statistical control and scale free topology.

In particular, **Figure 8** illustrates in stark contrast the gene-gene relationships which are found by analyzing identical datasets from different perspectives. All edge-sets depicted therein contain biological information, and each relies on different biological ground. No approach depicted is an incorrect analysis of the dataset described here, merely incomplete. A complete approach to an experiment with multiple datasets must analyze the data in parts, in comparison, and as a whole, in order to gain a complete picture of the behavior of those datasets when analyzed from multiple perspectives.

Especially notable is that our method tracks and identifies statistically significant changes in gene-gene interactions. This approach will allow for the tracking of these changes over a great many datasets. In many experimental systems – and some crops – there exists a great many microarray and high-throughput sequencing gene expression datasets. These datasets are deposited in massive centralized data warehouses, and infrequently accessed or re-analyzed. Integration of these data, along with other orthologous data, such as quantitative trait loci, exacting phenotypic analysis, and protein-protein interaction maps will allow for the construction of large scale models of plant functions and systems. Requisite for such an effort is a detailed understanding of gene-gene relationships, and how such relationships respond to changing environmental conditions, nutrient availability, tissues, and developmental stage.

Analysis of disparate experiments via direct network comparison – a proof of concept

4.1 Introduction

Over the last several years, the use of GCNA to characterize patterns of expression among genes or transcripts within a transcriptome has become increasingly popular. GCNs have been used to characterize the behavior of genes on a transcriptome-wide scale in plants under abiotic stress [69,135,136], biotic stresses [137–139], gene expression perturbations (for example, through TDNA insertional mutagenesis) [140], and cosmic rays [141]. Extensions to the standard application of GCNA to expression datasets have been completed by the addition of metabolomic datasets [142,143], and by overlaying the results of quantitative trait locus (QTL) analysis onto the resultant network [144]. This 'layering' of multiple datasets has been exceedingly fruitful, and will continue to be so.

However, the above referenced work, and the similar multitude of studies are nearly all concerned with the analysis of one or a few individual datasets. The analysis of multiple datasets, especially when sourced from multiple independent biological experiments, has proven to be a hurdle, which has only recently become an active area of research. The promise of GCNA, or any unsupervised learning/clustering approach is to simplify the analysis of many-sample datasets in a way that pairwise comparison of sample sets cannot. Unfortunately, the comparison of datasets in the manner that network comparison makes possible simply raises the previous limitation of pairwise comparisons to a new data scale. The integration of many datasets into a holistic gene co-expression network was incompletely explored in 2013 [115].

4.1.1 Analysis of very large-scale transcriptome datasets

Feltus et al, analyzed more than 7,100 publicly available *Arabidopsis* microarray gene expression datasets. Through a data mining approach which utilized pre-clustering of the array set into 86 non-overlapping groups, the authors were able to identify an unprecedented level of gene-gene interactions. When analyzed as a single, overarching dataset, the 7100 microarrays resulted in a network of only 3,297 total genes, and 129,134 interactions. In comparison, when the 86 groups of microarrays were analyzed as independent sets, the total set of all networks contained 19,588 genes (94.7% of those on the *Arabidopsis* Affymetrix chip) and 558,022 gene-gene interactions. A great deal of characterization was performed on each individual network. Each individual network contained a number of modules, which could be further characterized by individual gene connectivity, module-level functional and promoter-content enrichment, and conformity to scale-free topology.

However, the authors stopped short of comparing the networks directly. It would be an important extension of their work to compare the network-to-network changes in gene-gene associations in a robust manner. The application of the dGCNA algorithm to this problem would enable tracking of gene-gene associations across all possible comparisons of tissue, developmental stage, environmental conditions, and stresses. Unlike approaches which use targeted datasets to identify putative loci critical for targeted functions (i.e., heat stress response), the robust comparison of many datasets would allow both the targeted characterization of gene-gene interactions, but also the determination of the variance of those interactions in a robust manner.

Here, we characterize gene-gene co-expression relationships within, and between, fourteen microarray datasets describing gene expression of the model organism *Brachypodium distachyon* under abiotic stress, control, and coordinated circadian growth conditions. We track the co-expression relationships between genes, and determine which relationships change in a statistically significant fashion between conditions.

4.1.2 Author Contributions

Henry Priest conceived and executed all analyses, and wrote the manuscript.

4.2 Results

Each of the individual datasets, described in **Table 8,** was analyzed in an identical way. All datasets were compared in an all-versus-all fashion, in which parameters for the sigmoid adjacency function of dGCNA were determined, and permutation analysis was performed to identify the correct differential adjacency cutoff to utilize for determination of the final elasticity network structure. Finally, modular analysis via functional enrichment and promoter analysis for the enrichment of short sequences within annotated promoter regions was performed.

Table 8. Description of Input Datasets.

Experiment	Sampling Start (ZT)	Intervals (hours)	Temperature (C°)	Day Length (hours)
High Temp.	ZT+2	1, 2, 5, 10, 24	42	16
High Salinity	ZT+2	1, 2, 5, 10, 24	22	16
Drought	ZT+2	1, 2, 5, 10, 24	22	16
Chilling	ZT+2	1, 2, 5, 10, 24	4	16
Control A	ZT+2	1, 2, 5, 10, 24	22	16
Control B	ZT+2	1, 2, 5, 10, 24	22	16
Control C	ZT+2	1, 2, 5, 10, 24	22	16
LDHC	ZT+0	every 4 hours	28/12	12
LDHH	ZT+0	every 4 hours	28/28	12
LLHC	ZT+0	every 4 hours	28/12	12
LDHC Freerun	ZT+0	every 4 hours	28	24
LDHH Freerun	ZT+0	every 4 hours	28	24
LLHC Freerun	ZT+0	every 4 hours	28	24

4.2.1 Parameter and Cutoff Calculation

The appropriate parameters for use in the sigmoid adjacency function were calculated utilizing the ‘scalefree’ command within dGCNA. This represented 156 sets of parameters (alpha and mu) for the adjacency function for all possible comparisons of the thirteen input datasets (one parameter set is generated for each of the positive and negative elasticity networks in a single comparison). All scale free criteria and adjacency cutoffs utilized for network comparison are available in **http://www**.danforthcenter.org/hpriest/Supplemental_File_9.xlsx

Supplemental File ***10***.

Differential similarity cutoffs were found via the use of dGCNA's 'permute' command. This returned a set of 156 differential similarity cutoffs, which are summarized in **Figure 10**.

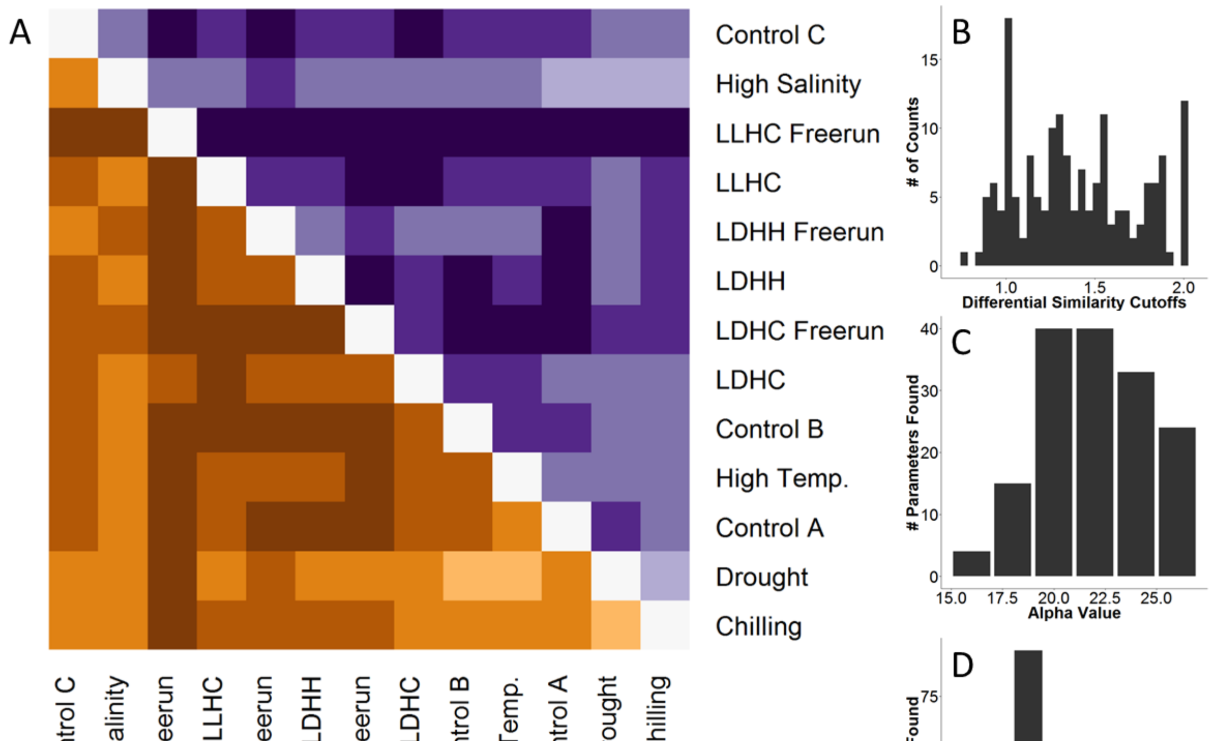


Figure 10. Elasticity Network Parameter and Cutoff Selections A) Heatmap of differential similarity cutoff values as determined by permutation of input datasets. Positive values (purples) represent cutoffs determined for positive elasticity networks. Negative values (oranges) represent cutoffs determined for negative elasticity networks. Diagonal values are zero, as networks were not compared to themselves. **B)** Histogram of the absolute values of all differential similarity cutoffs found by dGCNA 'permute' command. **C)** Histogram of values determined by dGCNA scalefree command for the alpha parameter of the sigmoid adjacency function. All even values on the interval [16,26] inclusive were interrogated. **D)** Histogram of values determined by dGCNA scalefree command for the alpha parameter of the sigmoid adjacency function. All values evenly divisible by 0.05 on the interval [0.70,0.95] inclusive were interrogated.

We observed distinct trends in the cutoffs determined for each comparison (**Figure 10A**). Most obviously, the LLHC Free-run (LLHC-FR) dataset has very high cutoffs in almost all comparisons. Indeed, in all but one comparison (against the High Salinity dataset), the cutoffs for significant elasticity exceeded a differential similarity of 1.5, and resulted in very small networks (see below). Comparisons against three abiotic stress datasets, Drought, Chilling, and High Salinity consistently yielded the lowest cutoffs. Lower cutoffs are naturally associated with higher network sizes, but are also the result of higher overall differences in gene-gene similarity scores on a network-wide scale.

Cutoffs determined by permutation followed a roughly normal distribution, with mean 1.43 (**Figure 10B**). A number of cutoffs were determined to be at positive or negative 2, indicating that there was no difference between the networks compared that exceeded that which would be expected by chance. This also results in an empty network, and in every case in which no significant differential similarities could be found, the LLHC-FR dataset was a member of the comparison.

Parameters utilized in the sigmoid function were determined by adherence of the produced network to the scale free criteria. Even in cases in which no significant differential similarities were found, scale free topology could be achieved in the elasticity networks of the datasets being compared. This indicates a possible pitfall of relying solely on scale-free topology in network construction, at least in the determination of differential network topology. Values for the alpha parameter were evenly distributed around the center of the range tested (**Figure 10C)**. However, values for the mu parameter were heavily biased toward the lower end of the range tested (**Figure 10D**). This is a product of utilizing the lowest parameter found to generate acceptable scale-free topology.

The cutoffs determined by permutation for statistical significance are identified in the context of differential similarity – prior to adjacency transformation by parameters determined to generate scale-free topology. Although in previous analyses it appeared that there was a strong relationship, in an elasticity context, between scale-free topology and statistically significant edges, in the analysis of these datasets, 36/156 adjacency-transformed cutoffs were non-zero, but on the interval [-0.1,0.1], indicating that scale free topology criteria identified adjacency parameters which were very close to those identified by permutation analysis. A further 10 cutoffs were found to be equal to zero after adjacency transformation, indicating that the adjacency transform was a more strict determination of edge presence than permutation analysis. 26 additional adjacency-transformed cutoffs were found, whose absolute values, A, satisfied $1.9 < A < 2$. These cutoffs are very high, which indicates that the parameters for the sigmoid adjacency function were much more lenient than the cutoffs determined by permutation. There does not appear to be a strong relationship, in comparisons between the 13 datasets analyzed here, between statistical significance of differential similarity, and scale-free topology generated by a sigmoid adjacency function.

4.2.2 Distribution of Network Sizes

A strong inverse correlation (r^2 of 0.64) between the value of differential adjacency applied, and the size of the resultant network. In other words, as cutoffs closer and closer to 2, or -2, are required for statistical significance, the networks grow smaller and smaller, as expected.

When broken into quartiles by a ranking of the number of edges present in the calculated total elasticity network, clear trends appear. The bottom quartile of networks (containing 20 total networks) identified by application of significance cutoffs contained very few edges. The largest network in this group comprised only 35,844 total edges. Seven of these 20 networks contained no edges at all in one or both of the positive- or negative-elasticity subnetworks. This group also contained ten of thirteen possible comparisons of the LLHC Free-run (LLHC-FR) dataset. Six of the remaining 10 non-LLHC-FR comparisons involved comparisons between stress-experiment control data and Diurnal/Circadian experiments, or between replicates of stress-experiment control data. Four comparisons in this lowest quartile had very few edges, but did not have a readily available explanation: High Temperature stress against LDHC-FR (12 total edges), LDHC against LLHC (27 total edges), LDHC-FR against LDHH (77 total edges), and LDHC-FR against LLHC (31,805 total edges).

The top quartile represents twenty very large networks, ranging from 17.1 million edges (Drought vs LDHC-FR) to 55.2 million edges (Drought vs Control B). Of these, seventeen networks involve a comparison against the High-Salinity or Drought datasets.

In order to determine if the apparent disparity in edge-set sizes in elasticity networks relating to particular datasets were distributed in a non-random way, each group of twelve total elasticity networks associated with each dataset were analyzed via the Wilcoxon signed-rank test. Five of thirteen datasets tended to give rise to elasticity networks that were significantly different than the expectation: Drought, High-Salinity, and Cold (maximum p-value, $p < 0.035$). These three datasets gave rise to networks larger than expected. The LDHH-FR dataset also tended to generate large elasticity networks (p-value < 0.0069). All other datasets gave rise to elasticity networks whose size did not differ significantly from the median, except for LLHC-FR, which gave rise to exceedingly small networks (p-value < 0.0031).

All significant elastic edge set sizes, positive, negative, and total, are available in **Supplemental File 11**.

4.2.3 Probe-set Stability Analysis

We wished to determine if we could utilize the connectivity of individual probe-sets as a metric to determine 'stability' of those probe-sets in multiple pairwise comparisons between network-scale datasets. To this end, we calculated the connectivity of all genes in all pairwise comparisons. Within the dGCNA algorithm, the parameters identified for the sigmoid adjacency function are selected based on their generation of a scale-free topology in the resultant positive and negative elasticity networks. Therefore, by design, most genes have few connections, and few genes have many connections. Within each pairwise comparison, all individual probe-sets were ranked to account for the highly non-normal distribution of probe-set connectivity scores. These ranks enable us to utilized the rank-percentile of each probeset's connectivity as a connectivity score.

We applied this method of probe-set scoring to sets of genes identified via individual network analysis to be enriched for particular functions of interest. We wished to determine if taking a multiple-comparison approach enriched the biological understanding or information underlying a particular grouping of genes.

In our previous analysis of abiotic stress response in *Brachypodium*, a cluster of genes ('Module 02', **Figure 4** and **Figure 5**), was down-regulated in response to abiotic stress. By over-enrichment analysis of Gene-Ontology terms, this module was found to be substantially enriched for many functions related to cell cycle and DNA replication, as well as cell wall biogenesis, growth, and metabolism. Matos et al., later found that by removing thermocycles from the growth regimen entraining the circadian clock, four week old *Brachypodium* plants no longer displayed rhythmic growth [145].

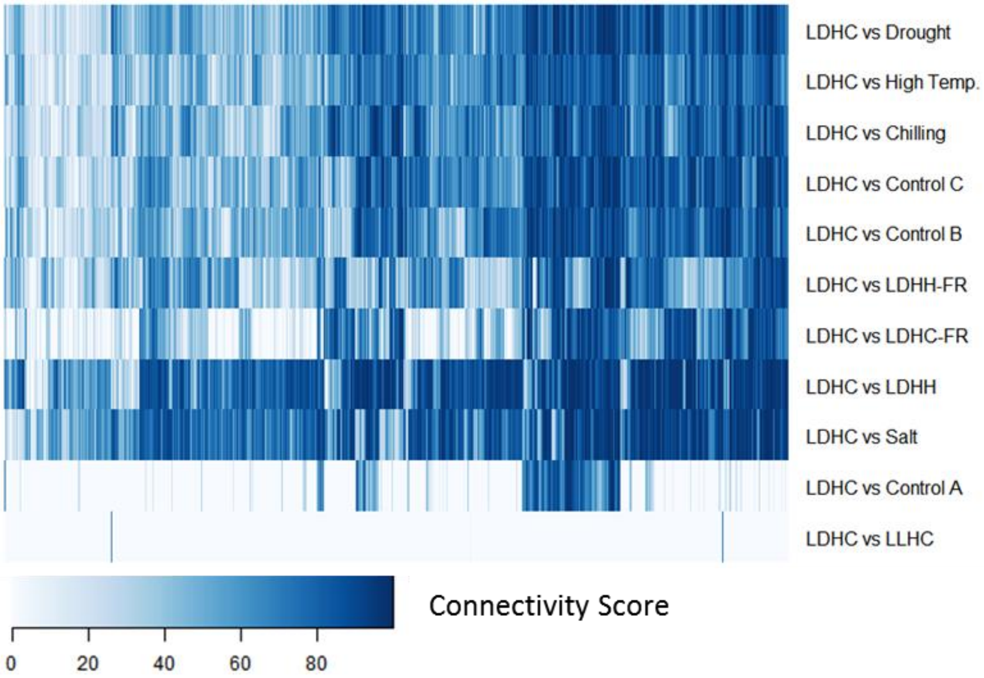


Figure 11. Gene-set Stability Analysis. Heatmap of per-probe-set rank scores for genes within Module 02 of abiotic stress response modular analysis. Module 02 is significantly enriched for genes related to DNA replication, cell wall biogenesis, the cell cycle, carbon metabolism, and growth. Comparison of LDHC and LDHH networks identified very high levels of elasticity in probe-sets aligning to genes of Module 02. Comparisons against LLHC, LLHC-FR, Control C not shown due to small network sizes (<120k edges, or average 4 connections/node).

We identified the probe-sets relating to set of genes contained in Module 02, and plotted their connectivity scores in elasticity networks derived from comparisons of the fully driven diurnal dataset against all other datasets (**Figure 11A**). A substantial portion of genes displayed very high elasticity in the comparison of LDHC against the LDHH dataset, which lacked thermocycles, and the high-salinity stress dataset. Of the total 976 gene loci present in Module 02, 878 gene loci were associated with at least one probe-set which demonstrated any elasticity between the datasets at all. Of these, 437 loci were in the top 15%, in terms of connectivity score, in the LDHC vs LDHH comparison, and/or the LDHC vs high salinity stress comparison. These genes comprise 44% of the original Module 02, and are enriched for largely the same GO terms as the original analysis identified. Of these high-elasticity genes, 182 (41.6%, 182/437) were found to be in the top 15%

of connectivity scores only in either the LDHC vs LDHH comparison, or the LDHC vs high-salinity stress comparison, but no other comparison. These genes were found to be enriched for GO terms relating exclusively to DNA replication and organization, microtubules, and the cell cycle. The complete list of all GO terms found to be statistically enriched in this set of genes is included in Error! Not a valid bookmark self-reference.. The promoters of these gene loci were enriched for forty different purine-rich elements (i.e., ‘AGAG’, and variants thereof). Despite the obvious diurnal expression pattern (**Figure 3A**), there was no enrichment for sequence motifs related to or resembling known circadian regulatory elements.

Table 9. Statistically enriched Gene Ontology terms of stress-responsive genes which experience elasticity under loss of thermocycles, or high salinity stress.

Gene Ontology Term	P-value	FDR-corrected P-value	Description
GO:0007018	6.40E-08	0.00010	microtubule-based movement
GO:0006928	7.20E-08	0.00010	cellular component movement
GO:0007017	1.90E-07	0.00017	microtubule-based process
GO:0034622	8.90E-06	0.00445	cellular macromolecular complex assembly
GO:0065003	9.80E-06	0.00445	macromolecular complex assembly
GO:0044085	1.60E-05	0.00545	cellular component biogenesis
GO:0006259	2.00E-05	0.00570	DNA metabolic process
GO:0034621	2.10E-05	0.00570	cellular macromolecular complex subunit organization
GO:0043933	2.30E-05	0.00570	macromolecular complex subunit organization
GO:0006270	4.60E-05	0.00964	DNA replication initiation
GO:0022607	7.40E-05	0.01344	cellular component assembly
GO:0016043	0.00017	0.02894	cellular component organization
GO:0006996	0.00021	0.03365	organelle organization
GO:0031497	0.00029	0.03838	chromatin assembly
GO:0034728	0.00029	0.03838	nucleosome organization
GO:0006334	0.00029	0.03838	nucleosome assembly
GO:0065004	0.00031	0.03838	protein-DNA complex assembly
GO:0006333	0.00035	0.03973	chromatin assembly or disassembly
GO:0009834	0.00035	0.03973	secondary cell wall biogenesis
GO:0006323	0.00038	0.04140	DNA packaging
GO:0051258	0.0004	0.04191	protein polymerization
GO:0007049	0.00048	0.04843	cell cycle

In addition, this grouping of elastic genes were characterized by a unique profile of expression. In visualizing the expression profile of these genes in both the LDHC and LDHH conditions, a severe shift in expression pattern is apparent (**Figure 12**). Under normal driven conditions (LDHC, **Figure 12A**), the genes in question display a characteristic diurnal expression pattern, reaching their maximum expression peak in the four to eight hours preceding dawn. These genes undergo positive elasticity, and a general un-coordination in the gene set's expression profile can also be seen in **Figure 12A**. Once thermocycles are removed (LDHH, **Figure 12B**), the gene set's expression pattern becomes both highly correlated and also strongly a-typical. In many cases of mis-regulation or mis-expression of circadian factors, the circadian oscillator rapidly damps until no oscillation can be observed [128,146]. Alternatively, small increases or decreases in the period of oscillation are also observed [147]. In this case, rapid changes in expression is observed among

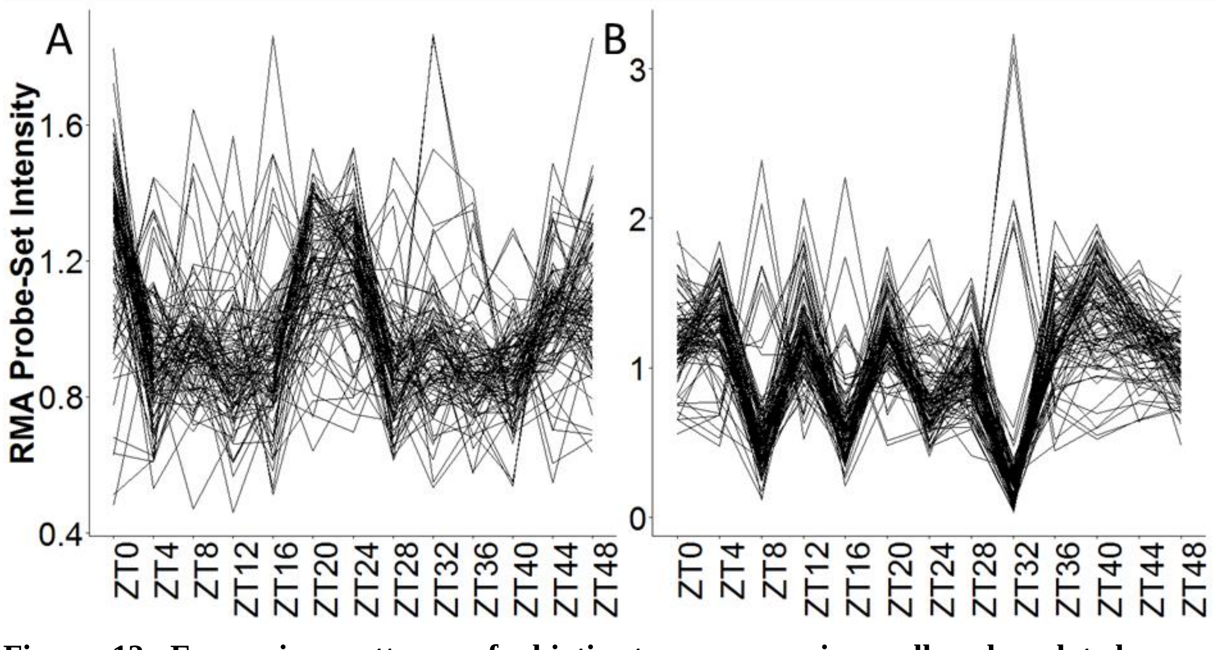


Figure 12. Expression patterns of abiotic-stress responsive, cell-cycle related genes experiencing positive elasticity under loss of thermocycles. A) Genes display a roughly diurnal expression pattern under normal, LDHC conditions. **B)** The same set of genes is expressed in a highly coordinated but atypical fashion under absence of thermocycles. Expression values are mean-normalized non-log RMA intensity values. ZT numbers on x-axis denote zeitgeber time, number of hours since onset of dawn. All light periods are twelve hours in length.

this set of high elasticity genes, in which the genes oscillate between peak and trough expression levels over repeated periods of four hours.

4.2.4 Prediction of Transcription Factor Binding Substrates

We wished to determine if network topology, or differential network topology, could be utilized to predict the interaction of DNA-binding proteins with their target substrates. We first theorized that transcription factors which are co-modular with other genes might be predicted to bind the promoters of those genes. The five transcription factors present in Module 07 of the original abiotic stress network were assayed in an all-versus-all approach against the promoters of five of the top ten highest connectivity gene loci in the same module. In two separate experimental replicates, 14 of the 25 tested interactions gave statistically significant increases in luciferase activity in the presence of the transcription-factor containing bait construct (**Figure 13**). In order to determine if these interactions could be predicted by differential network analysis, each elasticity network was mined for the all 25 of the possible edges depicted in Figure 4. The edge-values obtained were segregated into two groups, those that represented a positive interaction result from the yeast 1-hybrid assays, and those that represented a negative result. The Wilcoxon ranked sum test was utilized to determine if the two populations of edge values differed greatly. Two possible hypotheses were identified. Either, by virtue of their interaction on the molecular level, the transcription factor-target gene interactions would be more stable (i.e., lower overall elasticity), or, the ability of the transcription factor to bind the target promoters under certain conditions would cause the interactions to be more elastic. We therefore performed a two-way test, in which the null hypothesis is that the two populations of edge strengths do not differ, and the alternative hypothesis is that the two populations differ. This test returned a p-value of 0.034, and analysis of the populations indicates that the strengths of the positive interactions had higher elasticity values than those of the negative interactions.

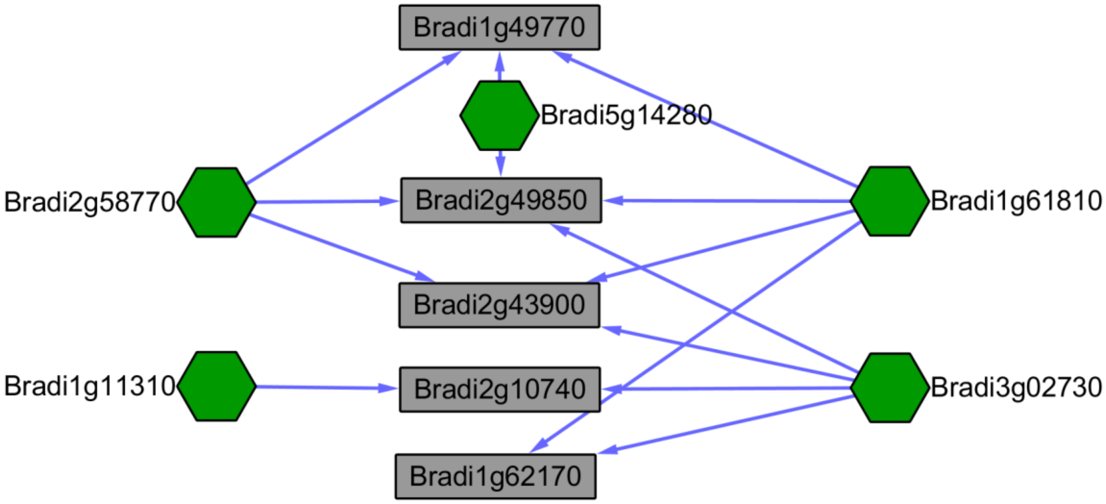


Figure 13. Interaction network of tested transcription factors and putative target promoters. Each interaction between each transcription factor (green hexagons) and putative target promoters (grey rectangles) are supported by two independent experimental replicates. Elasticity network interactions of the edges represented by blue arrows have higher elasticity than null interactions (non-connected TF/target gene pairs, p-value < 0.034, MWW test).

4.3 Discussion

Gene expression analysis has progressed through several complete transformations. Now, a wealth of gene expression data is available publicly on a multitude of species. The analysis of this data, on a large scale is a difficult challenge. We have taken an approach which relies on prior knowledge of the conditions that generated the datasets in question, and have conducted an *in-silico* experiment to determine if a statistically robust network comparison approach will yield informative results as to the datasets themselves, the gene loci and transcripts they represent, the relationships between those loci, and how those relationships depend on the conditions the organism inhabits.

4.3.1 Elasticity networks differ greatly in size

Differential network analysis via dGCNA enabled a detailed understanding of the datasets, and the networks they generate, as entire entities to themselves. Although all individual arrays and datasets passed quality control and filtering checks (see methods), it is quite clear that the biological experiment underlying LLHC-FR had serious underlying issues. It is probable at this point the issues lay in the execution of the experiment, as each step of RNA extraction, preparation, array hybridization, and quality control of arrays did not show any problems. While a conclusion may have been drawn in the context of the five other circadian experiments that the lack of light cycles combined with the free-run conditions led to no rhythmic gene expression (as prior analyses, not shown here, have indicated), this analysis indicates that no significant differences can be established with any other datasets consisting of an entirely different underlying structure. This probably indicates an underlying problem with the dataset, rather than a significant biological signal of interest.

An alternative interpretation of very small network sizes in a given pairwise comparison is that the two networks are highly similar in structure. In the extreme case, if a comparison between two identical networks were made, a null network would result. Thus, small networks are not only a product of underlying problems in the dataset. Distinguishing this from the above can be done by comparisons against many different datasets (as done here) or through construction of an individual network utilizing a single dataset (i.e., to determine if co-expression signals exist therein).

The number of identified significant elastic network edges differed greatly between individual comparisons. In general, this can be interpreted to represent the strength of the underlying biological differences between the experiments in question. Seventeen of the top twenty comparisons, in terms of network size, involved a stress condition, in each case either high-salinity or drought. Each of these abiotic stress conditions, in the original published analysis, identified a large number of differentially expressed genes, with more than 9000 genes differentially expressed under drought stress, nearly a third of all annotated genes in the *Brachypodium* genome. However, although the drought vs. control B comparison yielded a network of over 50 million edges, the control A vs control B comparison yielded a network of over one million edges. Despite this huge difference in size, this indicates substantial underlying changes in gene co-expression between control experiments.

Better than sample to sample correlations, and better than sample-correlation based clustering, comparisons of the emergent networks of array datasets allow for a fine-grained, edge-wise comparison of differences between datasets that generates an overall, dataset level comparison which illustrates the level of difference in gene relationships between large, sometimes unwieldy datasets.

4.3.2 Methods of identification of significant elasticity

The dGCNA java software supplies two methods for the identification of elasticity within a network comparison. In the original manuscript describing dGCNA, the cutoffs determined by permutation of input datasets, and by identification of scale-free topology were highly similar. It was of significant interest to determine if this phenomenon held true for the comparisons performed here. If scale-free elasticity networks also tended to consist of statistically significant edges, it would represent a convergence of statistics, biological signal, and network topography that is rarely found in computational science.

Naturally, as shown in **Figure 10**, there does not seem to be any major correlation between the networks generated by adjacency parameters which yield scale-free networks, and the networks generated by identification of statistically significant edges. Indeed, some networks had numerous edges present in a scale-free context, none of which were found to be statistically significant. Utilizing a generous classification of ‘similar’ cutoffs, in which the adjacency-transformed cutoff for statistical significance was less than 0.1 differential adjacency, only 46 of 156 total possible cutoffs were similar (29.4%).

4.3.3 Interaction of Disparate Datasets to Enrich Biological Signal

Originally identified in analysis of abiotic stress responses in *Brachypodium* (**Figure 3**), Module 02 is substantially enriched for genes relating to DNA replication, the cell cycle, DNA organization, growth, cell wall biogenesis, and carbon metabolism. The genes present in this module were compared by their connectivity scores. This analysis identified a large subset of genes which have significant elasticity between datasets derived from fully-driven, diurnal LDHC conditions, and from similar conditions which lack thermocycles (LDHH), as well as under high-salinity stress. By identifying 182 genes which undergo significant elasticity only between LDHC and LDHH, or LDHC and high-salinity (**Figure 11**), GO term enrichment analysis identified enrichment among terms relating only to DNA replication, organization, and the cell cycle. GO terms associated with growth and carbon metabolism were not present. The behavior of *Brachypodium* under loss of thermocycles has previously been studied. Matos et al. [145] found via high-throughput plant imaging analysis that rhythmic growth in *Brachypodium* was dependent primarily on the presence of thermocycles. The authors further concluded from microscopy that changes in growth rate were not due to either cell division or cell expansion changes alone.

In Handakumbura et al., it was found that in either of two loss-of-function-mutants for cellulose synthase genes (BdCESA4, BdCESA7), abnormal cell walls were generated, with less thick cell walls of both metaxylem and interfascicular fibers, as well as small stem area [148]. Both of these loci (Bradi3g28350 and Bradi4g30540, respectively) were associated with probesets that exhibited high elasticity in comparisons of abiotic stress datasets, as well as comparisons between stress and control datasets. Both loci were significantly more elastic in comparisons involving the LDHH dataset than they were in comparisons involving the LLHC dataset (Mann-Whitney-Wilcoxon, p-value<0.0006783 in both cases). It is evident that the incorporation and direct comparison of abiotic stress-related data to diurnal and circadian data yields strong biological signal.

An especially fascinating aspect of the set of 182 genes identified as highly elastic only between LDHC and LDHH, and between LDHC and high-salinity stress, is that these genes are decided not elastic between LDHC and drought stress. The responses of *Brachypodium distachyon* to drought and high-salinity stresses were highly similar overall (**Figure 5**). The genes captured as part of module 05 show strong down-regulation in response to drought, but a less strong response under high salinity. This could be taken to suggest that the response to drought seen in genes of module 05 does not decrease those genes' co-expression relationships that they possess under LDHC. However, the elasticity track of module 05 genes under high and low temperature stress is similar to that of drought, but their expression response under stress is more similar to that of high-salinity (**Figure 11**). There are very strong regulatory connections between the circadian clock, and abiotic stress responses in plants [129]. Given the major differences between the circadian system in *Brachypodium* (e.g., being much more strongly affected by thermocycles than by photocycles), the existence of links between various abiotic stresses should be examined more closely, and the analyses performed here are an excellent source of hypotheses.

Our analysis, which derives from a gene module identified to regulate cell growth and division under abiotic stress (conditions which generate small, sickly plants in the main), and enriched via the comparison of LDHC- and LDHH-derived datasets has identified a set of 182 genes. These genes are both responsive to stress and present in a cell-cycle enriched module, and they experience uniquely high elasticity in their gene-gene co-expression relationships upon loss of thermocycles (**Figure 11**). Further, this reduced gene set is not statistically enriched for functional terms relating to carbon metabolism or cell growth (We identified the probe-sets relating to set of genes contained in Module 02, and plotted their connectivity scores in elasticity networks derived from comparisons of the fully driven diurnal dataset against all other datasets (**Figure 11A**). A substantial portion of genes displayed very high elasticity in the comparison of LDHC against the LDHH dataset, which lacked thermocycles, and the high-salinity stress dataset. Of the total 976 gene loci present in Module 02, 878 gene loci were associated with at least one probe-set which demonstrated any elasticity between the datasets at all. Of these, 437 loci were in the top 15%, in terms of connectivity score, in the LDHC vs LDHH comparison, and/or the LDHC vs high salinity stress comparison. These genes comprise 44% of the original Module 02, and are enriched for largely the same GO terms as the original analysis identified. Of these high-elasticity genes, 182 (41.6%, 182/437) were found to be in the top 15% of connectivity scores only in either the LDHC vs LDHH comparison, or the LDHC vs high-salinity stress comparison, but no other comparison. These genes were found to be enriched for GO terms relating exclusively to DNA replication and organization, microtubules, and the cell cycle. The complete list of all GO terms found to be statistically enriched in this set of genes is included in Error! Not a valid bookmark self-reference.. The promoters of these gene loci were enriched for forty different purine-rich elements (i.e., ‘AGAG’, and variants thereof). Despite the obvious diurnal expression pattern (**Figure 3A**), there was no enrichment for sequence motifs related to or resembling known circadian regulatory elements.

Table 9). The expression patterns of these genes after loss of thermocycles is also highly atypical (**Figure 12B**), which indicates an interesting avenue of investigation may lie in the identification of regulatory motifs that drive such a pattern. The promoters analyzed here identified a large number of purine-rich sequences, which have been linked to transcript stability and transcriptional control [149,150]. Vaughn et al., in particular found that transcripts containing purine-rich elements in their 5' UTR region had a half-life of 9.6 hours, far greater than the average half-life of 3.8 hours. This temporal relation, in combination with the expression patterns of these genes under absence of thermocycles, which does not persist under loss of photocycles or entirely free-running conditions makes the regulatory circuitry around this gene set of great interest.

4.3.4 Relationship between network properties and molecular activity of transcription factor loci

Molecular interactions between transcription factors and their substrate DNA sequence motifs has been an area of active study for the last half-century [151]. The prediction of these interactions is a difficult problem in plant species, as plants as a whole are not highly amenable to transformation, can have large or complex genomes, and are utilized in research as whole organisms rather than in cell culture. These factors make the acquisition of ultra-high quality transcriptomic datasets a difficult task, which makes the application of rigorous computational learning or prediction techniques perfected in bacterial or cell culture systems difficult.

Development of high-throughput sequencing applications in gene expression profiling has enabled the rapid generation of expression data in a wide variety of plant systems, previously an impossible task utilizing microarray technology. Computational and bioinformatics methods appropriate to analyzing these datasets for the prediction of molecular interactions based on these types of data are an urgent need in plant science.

Here, five transcription factors were assayed for DNA-binding activity against 5 native promoter regions, all from *Brachypodium distachyon*. The transcription factor gene loci and the putative target promoter sequences were all drawn from the same gene module (Module 02, **Figure 3**). Yeast 1-Hybrid assays identified 14 interactions in two biological replicates (**Figure 13**). Analysis of the connectivity scores of the transcription factor loci did not reveal any results of particular interest. The transcription factors are not among those genes which experience extremely high elasticity, nor are they less elastic, than other gene loci on the whole. There was no statistically significant correlation between the connectivity scores of the putative binding targets and their cognate TFs across all pairwise comparisons of network data. Except in the same datasets that generated their co-modularity (i.e., high-salinity and drought against control experiments), there was no tendency for the gene loci to be co-modular. In short, there did not seem to be an enrichment on the network or modular levels for the positive molecular interactions, as verified by yeast 1-hybrid analysis.

However, analysis of edge-wise connectivity between TF loci and target promoters which interact on in the yeast 1-hybrid assay was significantly altered from those interactions which were found to not interact. The elasticity in adjacency scores between TF loci and their Y1H-identified promoter substrates was higher than those TF/target pairings found not to interact, at a statistically significant level (p-value < 0.034, MWW test). The signal found here is not particularly strong, and certainly does not indicate that TF/target gene elasticity can be used in exactly this manner, on a global transcriptome data level, to predict TF/target interactions. However, there may be an effect on TF/target elasticity for those pairs that truly interact *in vivo*. This certainly warrants further study with targeted expression datasets and tagged transcription factor proteins allowing verification of targets via immunoprecipitation followed by sequencing. This result, combined with that previously found in the analysis of an *Arabidopsis* LHY overexpression line (**Table 7**) strongly suggest molecular interactions between specific transcription factor and target gene loci may be an identifiable signal in whole-transcriptome datasets.

4.4 Conclusion

Comparative analysis of network data structures carry strong biological signals, even when comparing datasets from disparate experiments and conditions. However, final hypotheses that result from this type of analysis are still highly dependent on underlying experimental data, execution of experiments, biological rationales, and sources of data. No over-arching signals or hypotheses could be easily drawn without heavy reliance on biological understanding of regulatory systems and previous biological results. Node connectivity remains an informative measure of node importance. Signals within an all-by-all pairwise elasticity analysis context could be discerned on a whole network, gene-group (i.e., modular), and single-gene level. More development and experimentation must be conducted to determine if some detected biological signals are truly indicative of molecular function.

Fine-grained comparison of large-scale data sets of the nature executed here will become more and more necessary, as data is generated at an ever increasing rate. To not incorporate all available data across a particular plant system is a waste, of both the original data and the effort to warehouse the datasets in a rigorous manner. The well-reasoned and well-executed integration of a large number of disparate datasets can provide insight into gene function, and the variation of gene-to-gene relatedness across a multitude of environmental conditions, tissues, developmental stages, and circadian cycles.

4.5 Methods

4.5.1 Plant Growth, Microarray Dataset Generation and Normalization

Microarray datasets describing gene expression in *Brachypodium distachyon* were obtained from two accessions available in http://www.plexdb.org/: BD1, and BD2. Affymetrix Microarrays (chip BradiAR1b520742) were quality controlled utilizing the processes described in [98]. Collectively, the quality-controlled dataset comprises thirteen individual experiments – one entire experiment (high-light stress) was discarded due to pervasive quality-control issues. A single microarray in the heat-stress experiment was also discarded.

The BD2 dataset contains 104 microarrays grouped into 7 experiments, which describe independent applications of heat, drought, high salinity, and chilling stress, along with three independent control experiments, consisting of un-treated wild-type. These experiments were sampled in an asymmetric time-course design, with time-points at 1-, 2-, 5-, 10- and 24-hours after onset of stress conditions. Each time-point was sampled in biological triplicate [69].

The BD1 dataset contains 78 microarrays, grouped into six experiments. These microarrays collectively describe a set of circadian/diurnal time-courses. *Brachypodium distachyon* Bd21-0 plants were grown under photo/thermo-cycles (LDHC), photo-cycles (LDHH), or thermocycles alone (LLHC), for twenty one days.

Over a forty-eight hour period, whole aerial tissues of three week old individual *Brachypodium distachyon* accession Bd21-0 plants were sampled every four hours. Three different light and temperature conditions were applied for at least one week prior to the beginning of sampling, to entrain the circadian system. These conditions consisted of thermo-cycles (LLHC, 12 hours 28°C, 12 hours 12°C, constant light), photo-cycles (LDHH, 12 hours light, 12 hours dark, constant 28°C) and photo/thermo-cycles (LDHC, 12 hours light with 28°C, 12 hours dark with 12°C). Light intensity was set to 1000 μmol $m^{-2}s^{-1}$. Relative humidity was set to 50%. All plants were grown in a Conviron PGR 15 growth chamber. After the driven experiments were complete, remaining plants were placed under constant conditions (LLHH, utilizing the same light and temperature regimen above), for 24 hours. Following this spacer, plants were sampled in exactly the same manner, to produce three circadian time-courses, which we refer to as "free-run", or LDHC-FR, LDHH-FR, and LLHC-FR. All time-courses consist of thirteen time-points.

Leaf tissue preparation and RNA extraction was performed as described in [97]. RNA preparation, hybridization, chip scanning and QC were performed as described [98]. Specifically, A GeneChip® Fluidics Station 450 was used for hybridization, and hybridized arrays were scanned utilizing a GeneChip® Scanner 3000. Quality control was performed utilizing the standard procedure described within the Affymetrix protocols (Affymetrix GeneChip® Expression Analysis Technical Manual, 701021 Rev. 5; http://www.affymetrix.com). All molecular work was performed within the Oregon State University Center for Genome Research and Bioinformatics, Central Service Laboratory.

All microarray datasets included here were normalized as a set, utilizing the Robust Multi-array Average algorithm [70]. This algorithm was implemented in the Affymetrix Power Tools software package (http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx). The apt-probeset-summarize tool, version 1.15.0, also from Affymetrix, was used to summarize expression for each probeset.

4.5.2 Probeset inclusion criteria

All probesets were included, provided they satisfy the following criteria. In at least one of the thirteen included datasets, the probeset must have 80% of its total expression values exceeding a log2 RMA value of 7. This cutoff is explicitly chosen to be exceedingly lenient to allow probesets which are rarely expressed to be included, even if they are only expressed in a single experiment. The use of a comparative approach, statistical testing for differential similarity, and application of network topology criterion will limit nonsense signal created by inclusion of non-interesting probesets in certain datasets. These criteria resulted in an overall gene expression dataset of 182 data points describing a total of 30,993 individual microarray probesets. Some probesets describe the sense, and anti-sense strand of gene models included in the *Brachypodium distachyon* version 1.2 genome. However, these gene models are based on predictions, and do not yet incorporate data from recently available, strand-specific libraries. We therefore keep all included sense and anti-sense probesets which are co-genic as separate objects, as either could correctly describe the behavior of the gene locus in question.

4.5.3 Network Comparison

Network comparison between each possible pairwise combination of the input datasets was also completed using the dGCNA algorithm. dGCNA applies the scale-free topology criterion described above to identify proper values for use within a sigmoid adjacency function to create differential gene co-expression networks. The dGCNA command ‘scalefree’ was utilized along with the parameters: ‘aL’ = 16, ‘aH’ = 26, ‘mL’ = 0.7, ‘mH’ = 0.95, to determine proper parameters for alpha and mu within the sigmoid adjacency function. For these network comparisons, the minimum mu and alpha which generate a network possessing a correlation criterion greater than 0.8 and a slope less than -0.8 was selected.

In addition, dGCNA provides the ability to determine statistically significant differential adjacencies via the use of permutations of the input dataset. The ‘permute’ command was utilized, with 20 permutations, using the sigmoid adjacency function parameters identified above, to determine the statistical significance of each differential adjacency identified. All differential adjacencies identified which did not have a permutation-estimated false discovery rate of 0.05 or better were masked.

Finally, for each comparison, the final comparison engine was run (using the ‘compare’ command), utilizing the cutoffs identified above.

4.5.4 Yeast 1-Hybrid Assays

Prey constructs containing the targeted promoter sequences were grown in large liquid cultures. YU (MATα, tryptophan selection) yeast were utilized to harbor the prey construct. Each cloned promoter construct contains the 1000nt directly upstream of either the annotated transcription start site, or start codon, if no transcription start site is annotated. The promoter sequence of interest is cloned directly upstream of a luciferase reporter gene. Bait constructs containing the CDS of the transcription factor of interest are transformed into YM4271 (MATa, uracil selection). Bait and prey construct-containing yeast strains were co-incubated for 24 hours, incubated on diploid growth media (double selection, -UW) for 24 hours, and finally grown on complete media for a further 24 hours. Cultures are re-suspended in PBS and assayed within a clear-bottomed 384 well plate, via addition of the luciferin substrate coelenterazine.

Each promoter-TF interaction was assayed in 96 replicates. Activity was normalized against an empty-vector control bait. A separate analysis was conducted utilizing a non-activating reporter gene construct, with identical results. Luciferase activities falling more than 1.5*IQR from the media were excluded, for null, TF-promoter, and non-activating interactions. The TF-promoter activities were compared against the null and non-activating promoter activities via the non-parametric Mann-Whitney-Wilcoxon test. Only those interactions which were replicated in two separate experimental replicates were identified as positive.

4.5.5 Promoter Analysis

Promoter content analysis was completed utilizing the Element software [21]. Element analyzes a set of input promoter sequences for over-represented short sequences (5-8 nucleotides in length), by comparing the observed occurrence of each such short sequence in the input set, against the background set of all promoters present in a given genome. All statistical values generated by Element are corrected for false discovery.

4.5.6 Gene Ontology Analysis

The ‘topGO’ R package, available via Bioconductor [125,126] was utilized to test gene groups for over-representation of gene ontology terms. All p-values generated were corrected for multiple comparisons utilizing the R core function ‘p.adjust’, utilizing the ‘holm’ method for false discovery correction.

4.6 Acknowledgements

HDP conceived and conducted the analysis. **HDP** wrote the manuscript and made TF/target nominations. **JG, MW, DO** cloned sequences, created constructs, transformed yeast, operated robotics and conducted the Y1H assays. **HDP** analyzed data.

Future Directions

The generation of transgenic plants is always a difficult process. Even in well-studied systems, plant transformation efficiencies are low, and turn-around times are long [152–154]. Further, most transgenic cassettes are inserted at random within the genome and many methods of gene knock-down or knock-out are imprecise, incomplete, or have off-target effects. Underlying genetic differences that govern transformation efficiency can vary so strongly that individual accessions within a species may at times be completely intractable targets of transformation. In addition, the generation time in plant systems is often measured in weeks, rather than hours or days, making the cultivation and propagation of plant systems a lengthy and expensive process. In short, perturbing plant systems in a directed way is difficult, which makes many of the approaches that have been so powerful in prokaryotic and simple eukaryotes impossible for plant species that are not research models [155,156].

These factors combine to constrain many plant science research programs to working within the naturally occurring genetic variation of plant populations. Plant species are incredibly diverse. For example, the C_4 photosynthesis system – a core metabolic system responsible for energy harvesting – has independently evolved more than 45 times [157]. The genomes of domesticated varieties of crop plants (i.e., Wheat, *Triticum aestivum,* an allohexaploid, AABBDD, or Strawberry, *Fragaria vesca*, an octoploid [159,160]) are much more complex than research models. In all but the most well developed model systems, these factors all drive plant computational biology, on the whole, to be an engine for high-quality predictive analysis.

5.1 Data Mining & Machine Learning

A wealth of data already exists about many plant species, even crops of high importance to agriculture. More than 9,000 individual samples exist in the Gene Expression Omnibus for *Zea* *mays*, and more than 6,200 exist for *Glycine max* [161]. These data are underutilized, however ample evidence has shown that they cannot be simply analyzed in a straightforward manner to produce any intelligible results (**Figure 8**, [134]). Further development of the methods depicted in Gibson et al., and in Section 3, in combination with more rigorous machine learning techniques will leverage existing datasets in agriculturally important crops to identify answers to important questions. The results shown in Section 4 rely on an all-pairwise comparison schema of GCNs and reveal that understanding of biological signal is gained thereby. However, this is a cumbersome method of analysis when extended to any number of datasets. While it is clear that the differences between edgewise co-expression values hold useful information, individual edge values can be compared among a great many networks without conducting individual pairwise analyses. This is akin to analyzing the variance of gene expression across multiple samples, rather than the generation of differential expression values between a pair of such samples. The assembly of a many-dataset, edge-wise co-expression variance matrix will enable the identification of gene-gene relationships which co-vary in response to stress. By identifying those relationships known to exist on the molecular level (for example, between TFs and genes whose transcription is driven by the TF's target promoter), a supervised learning approach may shed light on the relationship between co-expression variance and molecular interactions.

A developing plant is a complex mosaic of tissue types, regulatory networks, intercellular signaling, metabolic cycles and stress responses. Each plant cell is able to modulate the state of its gene expression infrastructure in response to all of the above environments and stimuli. Gene-gene relationships change in drastic ways constantly. The ability to predict the tissues in which a particular gene pair relationship will exist will greatly identify the potential for a targeted perturbation to have unintended side effects. For example, if a TF and a target gene are co-expressed in a multitude of tissues, it may be the case that manipulating the expression of that TF locus will also affect the TF's target gene in those tissues. If, however, the co-expression relationship of the TF and target gene is isolated to a single tissue type, there may be an expectation that the impact of perturbations on the TF's protein level may only affect that specific target in that cell type.

The targeted analysis of datasets through machine learning and/or unsupervised data mining techniques will allow for the establishment of a gene co-expression 'galaxy', or collection of condition-specific gene co-expression networks. The appreciation of the elasticity of gene-gene relationships between conditions, or between networks, will allow for the understanding of gene relationships on an extremely fine-grained level.

The identification of stable gene-gene relationships across conditional subsets conversely requires the identification of highly unstable gene-gene relationships. These may be critical to normal function of plant systems, but represent poor targets for manipulation at the genetic level. Collectively, the assignation of gene-gene relationships as stable and non-stable across developmental stage, tissue, and environmental condition will allow high-quality predictions to be made regarding effects of transgenic perturbations, approaching an overarching goal of plant science, predicting phenotype as a product of genetic, epigenetic and environmental effects.

5.2 Integration of Orthogonal Datasets

The integration of multiple data types has been proceeding in both plant and animal systems for some time. This integration, for example, of quantitative trait loci, gene co-expression networks, and protein-protein interaction maps have been successful to varying degrees [142–144,164–166]. However these approaches often do not take a detailed approach towards integration of these data into a cohesive model. AraNet, for example, provides many data-types alongside one another, and does not approach gene expression analysis in the most fine-grained way. QTL identification is, by its nature, dependent on a great many factors, including population structures, the experimental design, and the environment or geographic location in which the plants are grown [167]. Especially as plants with fewer genomic resources or historical molecular characterization are studied, a more direct integration of genetic variation must be developed. Predicted regulatory effects of QTLs found in TF or long non-coding RNAs must be attributed to observed single nucleotide polymorphisms or insertions and deletions.

High throughput phenotyping platforms of multiple kinds are becoming more and more widespread, either through the installation of large controlled environment imaging platforms, field-deployed gantry and/or drone systems, or cheap, home-made optical sensor arrays can capture longitudinal data from plants through their growing cycle with reduced human interaction (citations). These systems will allow for the digitization of plant phenotypes that were previously categorically notated. For example, biotic infections can be assessed by the severity of surface lesions and necrotic areas, rather than scored in binary or gross severity levels. Plant height, growth rate, and color can be recorded on very large numbers of plants with high accuracy and low labor. The integration of phenotypic data with co-expression networks – especially on the metabolic level – has been a developing area of GCNA for some time [163]. The wide array of phenotypic data afforded by the flourishing field of high-throughput phenotyping will greatly enrich the utility of GCNs.

In a similar manner to optical phenotyping, the size, and cost of small environmental sensors are decreasing rapidly. Present in the idea of ‘precision farming’ but also in the most state-of-the-art greenhouses, is the tracking of micro-climates in plant growth environments on a large scale.

Extremely high density temperature and humidity data, along with either well-tracked precipitation data or tightly-controlled watering will allow even higher accuracy modeling of plant environment and nutrient availability.

In short, the data deluge we have observed in the last ten years is only the beginning of the avalanche of data we will receive in the next ten years. The finely targeted comparison of gene co-expression networks is a necessary small step to generate organism-scale, data-driven functional models of plant systems critical to meet many global needs.

References

1. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, et al. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science. 2000;290: 2105–2110. doi:10.1126/science.290.5499.2105

2. Guo AY, Chen X, Gao G, Zhang H, Zhu QH, Liu XC, et al. PlantTFDB: A comprehensive plant transcription factor database. Nucleic Acids Res. 2008;36: 966–969. doi:10.1093/nar/gkm841

3. Ihn T, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, et al. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science (80-). 2014;799. doi:10.1126/science.1075090

4. Gao F, Foat BC, Bussemaker HJ. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics. 2004;5: 31. doi:10.1186/1471-2105-5-31

5. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2010;431: 99–104. doi:10.1038/nature02800.Transcriptional

6. Yarragudi A, Parfrey LW, Morse RH. Genome-wide analysis of transcriptional dependence and probable target sites for Abf1 and Rap1 in Saccharomyces cerevisiae. Nucleic Acids Res. 2007;35: 193–202. doi:10.1093/nar/gkl1059

7. Zhu C, Byers K, McCord R, Shi Z, Berger M, Newburger D, et al. High-resolution DNA binding specificity analysis of yeast transcription factors. Genome Res. 2009; 556–566. doi:10.1101/gr.090233.108

8. Ucar D, Beyer A, Parthasarathy S, Workman CT. Predicting functionality of protein-DNA interactions by integrating diverse evidence. Bioinformatics. 2009;25: i137–44. doi:10.1093/bioinformatics/btp213

9. Schlecht U, Erb I, Demougin P, Robine N, Borde V, van Nimwegen E, et al. Genome-wide Expression Profiling, In Vivo DNA Binding Analysis, and Probabilistic Motif Prediction Reveal Novel Abf1 Target Genes during Fermentation, Respiration, and Sporulation in Yeast. Mol Biol Cell. 2008;19: 308–317. doi:10.1091/mbc.E07

10. Rombauts S, Florquin K, Lescot M, Marchal K, Rouzé P, van de Peer Y. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol. 2003;132: 1162–76. doi:10.1104/pp.102.017715.nomes

11. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004;5: 276–287. doi:10.1038/nrg1315

12. Tompa M, Li N, Bailey TL, Church G, De Moor B, Eskin E, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005;23: 137–144. doi:10.1038/nbt1053

13. van Helden J, André B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol. 1998;281: 827–842. doi:10.1006/jmbi.1998.1947

14. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics. 2001;17: 1113–1122. doi:10.1093/bioinformatics/17.12.1113

15. Pavesi G, Mauri G, Pesole G. An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics. 2001;17 Suppl 1: S207–14. doi:10.1093/bioinformatics/17.suppl_1.S207

16. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouzé P, et al. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol. 2002;9: 447–464. doi:10.1089/10665270252935566

17. Hudson ME, Quail PH. Identification of Promoter Motifs Involved in the Network of Phytochrome A-Regulated Gene Expression by Combined Analysis of Genomic Sequence and Microarray Data. Plant Physiol. 2003;133: 1605–1616. doi:10.1104/pp.103.030437.DNA

18. Marino-Ramirez L. Statistical analysis of over-represented words in human promoter sequences. Nucleic Acids Res. 2004;32: 949–958. doi:10.1093/nar/gkh246

19. Nemhauser JL, Mockler TC, Chory J. Interdependency of Brassinosteroid and Auxin Signaling in Arabidopsis. PLoS Biol. 2004;2: e258. doi:10.1371/journal.pbio.0020258

20. Mockler TC, Michael TP, Priest HD. The Diurnal Project : Diurnal and Circadian Expression Profiling , Model-based Pattern Matching , and Promoter Analysis The Diurnal Project : Diurnal and Circadian Expression Profiling , Model-based Pattern Matching , and Promoter Analysis. 2007;LXXII: 353–363. doi:10.1101/sqb.2007.72.006

21. Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, et al. Network Discovery Pipeline Elucidates Conserved Time-of-Day–Specific cis-Regulatory Modules. Takahashi JS, editor. PLoS Genet. Public Library of Science; 2008;4: 17. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2222925&tool=pmcentrez&rendertype=abstract

22. Kreps J, Budworth P, Goff S, Wang R. Identification of putative plant cold responsive regulatory elements by gene expression profiling and a pattern enumeration algorithm. Plant Biotechnol J. 2003;1: 345–352. doi:10.1046/j.1467-7652.2003.00032.x

23. Toufighi K, Brady SM, Austin R, Ly E, Provart NJ. The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. Plant J. 2005;43: 153–163. doi:10.1111/j.1365-313X.2005.02437.x

24. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006;34: W369–W373. doi:10.1093/nar/gkl198

25. Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, et al. A gene expression map of the Arabidopsis root. Science. 2003;302: 1956–1960. doi:10.1126/science.1090022

26. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, et al. A gene expression map of Arabidopsis thaliana development. Nat Genet. 2005;37: 501–506. doi:10.1038/ng1543

27. Brady SM, Orlando DA, Lee J-Y, Wang JY, Koch J, Dinneny JR, et al. A High-Resolution Root Spatiotemporal Map Reveals Dominant Expression Patterns. Science (80-). 2007;318: 801–806.

28. Zhang W, Ruan J, Ho T -h. D, You Y, Yu T, Quatrano RS. Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana. Bioinformatics. 2005;21: 3074–3081. doi:10.1093/bioinformatics/bti490

29. Koussevitzky S, Nott A, Mockler TC, Hong F, Sachetto-Martins G, Surpin M, et al. Signals from chloroplasts converge to regulate nuclear gene expression. Science. 2007;316: 715–719. doi:10.1126/science. 1140516

30. Walley JW, Coughlan S, Hudson ME, Covington MF, Kaspi R, Banu G, et al. Mechanical Stress Induces Biotic and Abiotic Stress Responses via a Novel cis-Element. PLoS Genet. 2007;3: e172. doi:10.1371/journal.pgen.0030172

31. Evrard A, Ndatimana T, Eulgem T. FORCA, a promoter element that responds to crosstalk between defense and light signaling. BMC Plant Biol. 2009;9: 2. doi:10.1186/1471-2229-9-2

32. Covington MF, Maloof JN, Straume M, Kay SA, Harmer SL. Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. Genome Biol. 2008;9: R130. doi:10.1186/gb-2008-9-8-r130

33. Michael TP, Breton G, Hazen SP, Priest H, Mockler TC, Kay SA, et al. A Morning-Specific Phytohormone Gene Expression Program underlying Rhythmic Plant Growth. Weigel D, editor. PLoS Biol. Public Library of Science; 2008;6: 12. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2535664&tool=pmcentrez&rendertype=abstract

34. Zdepski A, Wang W, Priest HD, Ali F, Alam M, Mockler TC, et al. Conserved Daily Transcriptional Programs in Carica papaya. Trop Plant Biol. Springer-Verlag; 2008;1: 236–245. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2890329&tool=pmcentrez&rendertype=abstract

35. Walther D, Brunnemann R, Selbig J. The regulatory code for transcriptional response diversity and its relation to genome structural properties in A. thaliana. PLoS Genet. 2007;3: e11. doi:10.1371/journal.pgen.0030011

36. Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y. Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. Plant Physiol. 2009;150: 535–546. doi:10.1104/pp.109.136028

37. Giuliano G, Pichersky E, Malik VS, Timko MP, Scolnik P a, Cashmore a R. An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. Proc Natl Acad Sci U S A. 1988;85: 7089–93. doi:10.1073/pnas.85.19.7089

38. Freeling M, Rapaka L, Lyons E, Pedersen B, Thomas BC. G-boxes, bigfoot genes, and environmental response: characterization of intragenomic conserved noncoding sequences in Arabidopsis. Plant Cell. 2007;19: 1441–1457. doi:10.1105/tpc.107.050419

39. Freeling M, Subramaniam S. Conserved noncoding sequences (CNSs) in higher plants. Curr Opin Plant Biol. 2009;12: 126–132. doi:10.1016/j.pbi.2009.01.005

40. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science (80-). 2007;316: 1497–1503.

41. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448: 553–560. doi:10.1038/nature06008.Genome-wide

42. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. Dynamic Remodeling of Individual Nucleosomes Across a Eukaryotic Genome in Response to Transcriptional Perturbation. Rando OJ, editor. PLoS Biol. Public Library of Science; 2008;6: 13. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2267817&tool=pmcentrez&rendertype=abstract

43. ’t Hoen P a C, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, et al. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. 2008;36: e141. doi:10.1093/nar/gkn705

44. Pruneda-Paz JL, Breton G, Para A, Kay SA. A Functional Genomics Approach Reveals CHE as a Component of the Arabidopsis Circadian Clock. Science (80-). 2009;323: 202–209. doi:10.1007/s12374-009-9030-1

45. Priest HD, Filichkin SA, Mockler TC. Cis-regulatory elements in plant cell signaling. Curr Opin Plant Biol. Elsevier Ltd; 2009;12: 643–649. Available: http://www.ncbi.nlm.nih.gov/pubmed/19717332

46. Wang W, Vinocur B, Altman A. Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. Planta. 2003;218: 1–14. doi:10.1007/s00425-003-1105-5

47. Witcombe JR, Hollington PA, Howarth CJ, Reader S, Steele KA. Breeding for abiotic stresses for sustainable agriculture. Philos Trans R Soc Lond B Biol Sci. 2008;363: 703–16. doi:10.1098/rstb.2007.2179

48. Mahajan S, Tuteja N. Cold, salinity and drought stresses: an overview. Arch Biochem Biophys. 2005;444: 139–58. doi:10.1016/j.abb.2005.10.018

49. Hirayama T, Shinozaki K. Research on plant abiotic stress responses in the post-genome era: past, present and future. Plant J. 2010;61: 1041–52. doi:10.1111/j.1365-313X.2010.04124.x

50. Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, et al. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J. 2007;50: 347–63. doi:10.1111/j.1365-313X.2007.03052.x

51. Zhou J, Wang X, Jiao Y, Qin Y, Liu X, He K, et al. Global genome expression analysis of rice in response to drought and high-salinity stresses in shoot, flag leaf, and panicle. Plant Mol Biol. 2007;63: 591–608. doi:10.1007/s11103-006-9111-1

52. Matsui A, Ishida J, Morosawa T, Mochizuki Y, Kaminuma E, Endo T a, et al. Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. Plant Cell Physiol. 2008;49: 1135–49. doi:10.1093/pcp/pcn101

53. Zeller G, Henz SR, Widmer CK, Sachsenberg T, Rätsch G, Weigel D, et al. Stress-induced changes in the Arabidopsis thaliana transcriptome analyzed using whole-genome tiling arrays. Plant J. 2009;58: 1068–82. doi:10.1111/j.1365-313X.2009.03835.x

54. Araus JL. Plant Breeding and Drought in C3 Cereals: What Should We Breed For? Ann Bot. 2002;89: 925–940. doi:10.1093/aob/mcf049

55. Ashraf M, Hayat MQ, Jabeen S, Shaheen N, Ajab M, Yasmin G. Artemisia L . species recognized by the local community of northern areas of Pakistan as folk therapeutic plants. J Med Plants Res. 2010;4: 112–119.

56. Chew YH, Halliday KJ. A stress-free walk from Arabidopsis to crops. Curr Opin Biotechnol. Elsevier Ltd; 2011;22: 281–6. doi:10.1016/j.copbio.2010.11.011

57. Zhu J. Cell signaling under salt , water and cold stresses. Curr Opin Plant Biol. 2001;4: 401–406.

58. Rowley ER, Mockler TC. Plant Abiotic Stress : Insights from the Genomics Era. Abiotic Stress Response in Plants - Physiological, Biochemical and Genetic Perspectives. Intech Open Access Journals; 2011.

59. Parre E, Ghars MA, Leprince A-S, Thiery L, Lefebvre D, Bordenave M, et al. Calcium signaling via phospholipase C is essential for proline accumulation upon ionic but not nonionic hyperosmotic stresses in Arabidopsis. Plant Physiol. 2007;144: 503–12. doi:10.1104/pp.106.095281

60. Tuteja N. Mechanisms of high salinity tolerance in plants. Methods Enzymol. 2007;428: 419–38. doi:10.1016/S0076-6879(07)28024-3

61. Doherty CJ, Van Buskirk HA, Myers SJ, Thomashow MF. Roles for Arabidopsis

CAMTA transcription factors in cold-regulated gene expression and freezing tolerance. Plant Cell. 2009;21: 972–84. doi:10.1105/tpc.108.063958

62. Moellering ER, Muthan B, Benning C. Freezing tolerance in plants requires lipid remodeling at the outer chloroplast membrane. Science (80-). 2010;330: 226–8. doi:10.1126/science.1191803

63. Lopushinsky W. Stomatal Closure in Conifer Seedlings in Response to Leaf Moisture Stress. Univ Chicago Press. 1969;130: 258–263.

64. Oliveira G, Peñuelas J. Effects of winter cold stress on photosynthesis and photochemical efficiency of PSII of the Mediterranean Cistus albidus L . and Quercus ilex L . Plant Ecol. 2004;175: 179–191.

65. Brinker M, Brosché M, Vinocur B, Abo-Ogiala A, Fayyaz P, Janz D, et al. Linking the salt transcriptome with physiological responses of a salt-resistant Populus species as a strategy to identify genes important for stress acclimation. Plant Physiol. 2010;154: 1697–709. doi:10.1104/pp.110.164152

66. Yazaki J, Shimatani Z, Hashimoto A, Nagata Y, Fujii F, Kojima K, et al. Transcriptional profiling of genes responsive to abscisic acid and gibberellin in rice: phenotyping and comparative analysis between rice and Arabidopsis. Physiol Genomics. 2004;17: 87–100. doi:10.1152/physiolgenomics.00201.2003

67. Janská A, Aprile A, Zámečník J, Cattivelli L, Ovesná J. Transcriptional responses of winter barley to cold indicate nucleosome remodelling as a specific feature of crown tissues. Funct Integr Genomics. 2011;11: 307–25. doi:10.1007/s10142-011-0213-8

68. Winfield MO, Lu C, Wilson ID, Coghill JA, Edwards KJ. Plant responses to cold: Transcriptome analysis of wheat. Plant Biotechnol J. 2010;8: 749–71. doi:10.1111/j.1467-7652.2010.00536.x

69. Priest HD, Fox SE, Rowley ER, Murray JR, Michael TP, Mockler TC. Analysis of Global Gene Expression in Brachypodium distachyon Reveals Extensive Network Plasticity in Response to Abiotic Stress. Wu K, editor. PLoS One. 2014;9: e87499. doi:10.1371/journal.pone.0087499

70. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4: 249–64. doi:10.1093/biostatistics/4.2.249

71. Lauria M, Rossi V. Epigenetic control of gene regulation in plants. Biochim Biophys Acta. Elsevier B.V.; 2011;1809: 369–78. doi:10.1016/j.bbagrm.2011.03.002

72. Bai L, Morozov A V. Gene regulation by nucleosome positioning. TRENDS Genet. Elsevier Ltd; 2010;26: 476–83. doi:10.1016/j.tig.2010.08.003

73. Zhong L, Xu Y, Wang J. DNA-methylation changes induced by salt stress in wheat Triticum aestivum. African J Biotechnol. 2009;8: 6201–6207. Available:

http://www.ajol.info/index.php/ajb/article/view/66122

74. Mukhopadhyay P, Singla-pareek SL, Reddy MK, Sopory SK. Stress-Mediated Alterations in Chromatin Architecture Correlate with Down-Regulation of a Gene Encoding 60S rpL32 in Rice. Plant Cell Physiol. 2013;54: 528–540. doi:10.1093/pcp/pct012

75. Marino D, Froidure S, Canonne J, Ben Khaled S, Khafif M, Pouzet C, et al. Arabidopsis ubiquitin ligase MIEL1 mediates degradation of the transcription factor MYB30 weakening plant defence. Nat Commun. 2013;4: 1476. doi:10.1038/ncomms2479

76. Lindemose S, O'Shea C, Jensen MK, Skriver K. Structure, function and networks of transcription factors involved in abiotic stress responses. Int J Mol Sci. 2013;14: 5842–78. doi:10.3390/ijms14035842

77. Zdobnov EM, Apweiler R. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17: 847–848.

78. Salvucci ME, Osteryoung KW, Crafts-brandner SJ, Vierling E. Exceptional Sensitivity of Rubisco Activase to Thermal Denaturation in Vitro and in Vivo 1. Plant Physiol. 2001;127: 1053–1064. doi:10.1104/pp.010357.1

79. Aranjuelo I, Molero G, Erice G, Avice JC, Nogués S. Plant physiology and proteomics reveals the leaf response to drought in alfalfa (Medicago sativa L.). J Exp Bot. 2011;62: 111–23. doi:10.1093/jxb/erq249

80. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40: D1178–86. doi:10.1093/nar/gkr944

81. Fujita Y, Fujita M, Shinozaki K, Yamaguchi-Shinozaki K. ABA-mediated transcriptional regulation in response to osmotic stress in plants. J Plant Res. 2011;124: 509–25. doi:10.1007/s10265-011-0412-3

82. Kim K, Portis AR. Temperature dependence of photosynthesis in Arabidopsis plants with modifications in Rubisco activase and membrane fluidity. Plant Cell Physiol. 2005;46: 522–30. doi:10.1093/pcp/pci052

83. Flexas J, Diaz-Espejo A, Galmés J, Kaldenhoff R, Medrano H, Ribas-Carbo M. Rapid variations of mesophyll conductance in response to changes in CO2 concentration around leaves. Plant Cell Environ. 2007;30: 1284–98. doi:10.1111/j.1365-3040.2007.01700.x

84. Chaves MM, Flexas J, Pinheiro C. Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. Ann Bot. 2009;103: 551–60. doi:10.1093/aob/mcn125

85. Haga N, Kobayashi K, Suzuki T, Maeo K, Kubo M, Ohtani M, et al. Mutations in MYB3R1 and MYB3R4 cause pleiotropic developmental defects and preferential down-regulation of multiple G2/M-specific genes in Arabidopsis. Plant Physiol. 2011;157: 706–17. doi:10.1104/pp.111.180836

86. Knight H, Trewavas AJ, Knight MR. Cold calcium signaling in Arabidopsis involves two cellular pools and a change in calcium signature after acclimation. Plant Cell. 1996;8: 489–503. doi:10.1105/tpc.8.3.489

87. Cheong YH, Kim K, Pandey GK, Gupta R, Grant JJ, Luan S. CBL1, a Calcium Sensor That Differentially Regulates Salt, Drought, and Cold Responses in Arabidopsis. Plant Cell. 2003;15: 1833–1845. doi:10.1105/tpc.012393.genes

88. Liu H, Li B, Shang Z, Li X, Mu R, Sun D, et al. Calmodulin Is Involved in Heat Shock Signal Transduction in Wheat. Plant Physiol. 2003;132: 1186–1195. doi:10.1104/pp.102.018564.large

89. Zhao H-J, Tan J-F. Role of calcium ion in protection against heat and high irradiance stress-induced oxidative damage to photosynthesis of wheat leaves. Photosynthetica. 2005;43: 473–476. doi:10.1007/s11099-005-0076-0

90. Qin D, Wu H, Peng H, Yao Y, Ni Z, Li Z, et al. Heat stress-responsive transcriptome analysis in heat susceptible and tolerant wheat (Triticum aestivum L.) by using Wheat Genome Array. BMC Genomics. 2008;9: 432. doi:10.1186/1471-2164-9-432

91. Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, et al. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 2002;30: 325–7. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99092&tool=pmcentrez&rendertype=abstract

92. Yamaguchi-Shinozaki K, Shinozaki K. A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. Plant Cell. 1994;6: 251–64. doi:10.1105/tpc.6.2.251

93. Stockinger EJ, Gilmour SJ, Thomashow MF. Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. Proc Natl Acad Sci U S A. 1997;94: 1035–40. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=19635&tool=pmcentrez&rendertype=abstract

94. Lim H, Cho M-H, Jeon J-S, Bhoo SH, Kwon Y-K, Hahn T-R. Altered expression of pyrophosphate: fructose-6-phosphate 1-phosphotransferase affects the growth of transgenic Arabidopsis plants. Mol Cells. 2009;27: 641–9. doi:10.1007/s10059-009-0085-0

95. Turano FJ, Fang TK. Characterization of two glutamate decarboxylase cDNA clones from Arabidopsis. Plant Physiol. 1998;117: 1411–21. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=34905&tool=pmcentrez&rendertype=abstract

96. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. BioMed Central; 2008;9: 559. Available:

http://www.ncbi.nlm.nih.gov/pubmed/19114008

97. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, et al. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res. Cold Spring Harbor Laboratory Press; 2010;20: 45–58. Available: http://www.ncbi.nlm.nih.gov/pubmed/19858364

98. Li C, Rudi H, Stockinger EJ, Cheng H, Cao M, Fox SE, et al. Comparative analyses reveal potential uses of Brachypodium distachyon as a model for cold stress responses in temperate grasses. BMC Plant Biol. 2012;12: 65. doi:10.1186/1471-2229-12-65

99. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. Oxford University Press; 2009;25: 1754–1760. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract

100. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98: 5116–21. doi:10.1073/pnas.091062498

101. Le S, Josse J, Husson F. FactoMineR : An R Package for Multivariate Analysis. J Stat Softw. 2008;25: 1–18.

102. Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. 2010;38: W64–70. doi:10.1093/nar/gkq310

103. The International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature. 2010;463: 763–8. doi:10.1038/nature08747

104. Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LGG, Rensing SA, Kersten B, Mueller-Roeber B. PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res. 2010;38: D822–D827. doi:10.1093/nar/gkp805

105. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. J Stat Softw. 2012;46.

106. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27: 431–2. doi:10.1093/bioinformatics/btq675

107. Fisher RA, Yates F. Statistical Tables: For Biological, Agricultural and Medical Research. 6th ed. New York: Hafner Press; 1963.

108. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457: 551–6. doi:10.1038/nature07723

109. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. Nature Publishing Group; 2010;463: 178–83. doi:10.1038/nature08670

110. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. Oxford University Press; 2011;40: D1202–10. doi:10.1093/nar/gkr1090

111. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009;326: 1112–5. doi:10.1126/science.1178534

112. Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, et al. Reference genome sequence of the model plant Setaria. Nat Biotechnol. Nature Publishing Group; 2012;30: 555–61. doi:10.1038/nbt.2196

113. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. 2007;35: D883–7. doi:10.1093/nar/gkl976

114. Pearson K. Notes on the History of Correlation. Biometrika Trust. 1920;13: 25–45.

115. Feltus FA, Ficklin SP, Gibson SM, Smith MC. Maximizing capture of gene co-expression relationships through pre-clustering of input expression samples: an Arabidopsis case study. BMC Syst Biol. BMC Systems Biology; 2013;7: 44. doi:10.1186/1752-0509-7-44

116. Gill R, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics. 2010; Available: http://www.biomedcentral.com/1471-2105/11/95

117. Ma C, Xin M, Feldmann KA, Wang X. Machine Learning-Based Differential Network Analysis: A Study of Stress-Responsive Transcriptomes in Arabidopsis. Plant Cell. 2014;26: 520–537. doi:10.1105/tpc.113.121913

118. Gambardella G, Moretti MN, De Cegli R, Cardone L, Peron A, Di Bernardo D. Differential network analysis for the identification of condition-specific pathway activity and regulation. Bioinformatics. 2013;29: 1776–1785. doi:10.1093/bioinformatics/btt290

119. Fuller TF, Ghazalpour A, Aten JE, Drake T a., Lusis AJ, Horvath S. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome. 2007;18: 463–472. doi:10.1007/s00335-007-9043-3

120. Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics. BioMed Central Ltd; 2010;11: 497. doi:10.1186/1471-2105-11-497

121. Li A, Horvath S. Network neighborhood analysis with the multi-node topological overlap measure. Bioinformatics. 2007;23: 222–31. doi:10.1093/bioinformatics/btl581

122. Barabási a. L. Statistical mechanics of complex networks. Rev Mod Phys. 2002;74: 48–94. doi:10.1103/RevModPhys.74.47

123. Zhang B, Horvath S. A general framework for weighted gene co-expression network

analysis. Stat Appl Genet Mol Biol. 2005;4: Article17. doi:10.2202/1544-6115.1128

124. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics. 2012;13: 328. doi:10.1186/1471-2105-13-328

125. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006;22: 1600–1607. doi:10.1093/bioinformatics/btl140

126. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5: R80. doi:10.1186/gb-2004-5-10-r80

127. Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, et al. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genet. 2008;4: e14. doi:10.1371/journal.pgen.0040014

128. Schaffer R, Ramsay N, Samach A, Corden S, Putterill J, Carré I a., et al. The late elongated hypocotyl mutation of Arabidopsis disrupts circadian rhythms and the photoperiodic control of flowering. Cell. 1998;93: 1219–1229. doi:10.1016/S0092-8674(00)81465-8

129. Dong M a, Farré EM, Thomashow MF. Circadian clock-associated 1 and late elongated hypocotyl regulate expression of the C-repeat binding factor (CBF) pathway in Arabidopsis. Proc Natl Acad Sci U S A. 2011;108: 7241–6. doi:10.1073/pnas.1103741108

130. Harmer SL, Hogenesch JB, Straume M, Chang HS, Han B, Zhu T, et al. Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. Science. 2000;290: 2110–2113. doi:10.1126/science.290.5499.2110

131. Dowson-Day M, Millar A. Circadian dysfunction causes aberrant hypocotyl elongation patterns in Arabidopsis. Plant J. 1999;17: 63–71. doi:10.1046/j.1365-313X.1999.00353.x

132. Roden LC, Ingle R a. Lights, rhythms, infection: the role of light and the circadian clock in determining the outcome of plant-pathogen interactions. Plant Cell. 2009;21: 2546–2552. doi:10.1105/tpc.109.069922

133. Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics. BioMed Central Ltd; 2010;11: 497. doi:10.1186/1471-2105-11-497

134. Gibson SM, Ficklin SP, Isaacson S, Luo F, Feltus F a, Smith MC. Massive-scale gene co-expression network construction and robustness testing using random matrix theory. PLoS One. 2013;8: e55871. doi:10.1371/journal.pone.0055871

135. Sarkar NK, Kim Y-K, Grover A. Coexpression network analysis associated with call of rice seedlings for encountering heat stress. Plant Mol Biol. 2014;84: 125–43. doi:10.1007/s11103-013-0123-3

136. Gene P, Sharma A. In silico identi fi cation of regulatory motifs in co-expressed genes under osmotic stress representing their co-regulation. Plgene. Elsevier B.V.; 2015;1: 29–34. doi:10.1016/j.plgene.2015.01.001

137. Amrine KCH, Blanco-Ulate B, Cantu D. Discovery of core biotic stress responsive genes in Arabidopsis by weighted gene co-expression network analysis. PLoS One. 2015;10: e0118731. doi:10.1371/journal.pone.0118731

138. Zheng Z-L, Zhao Y. Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to "Candidatus Liberibacter asiaticus" infection. BMC Genomics. BMC Genomics; 2013;14: 27. doi:10.1186/1471-2164-14-27

139. Righetti K, Vu JL, Pelletier S, Vu BL, Glaab E, Lalanne D, et al. Inference of Longevity-Related Genes from a Robust Coexpression Network of Seed Maturation Identifies Regulators Linking Seed Storability to Biotic Defense-Related Pathways. Plant Cell. 2015;27: tpc.15.00632. doi:10.1105/tpc.15.00632

140. Thomson B, Raganelli A, Wuest SE, Ryan PT, Maoil DSO, Kwa K, et al. Gene network analysis of Arabidopsis thaliana flower development through dynamic gene perturbations. 2015; 344–358. doi:10.1111/tpj.12878

141. Hwang S-G, Kim DS, Hwang JE, Han A-R, Jang CS. Identification of rice genes associated with cosmic-ray response via co-expression gene network analysis. Gene. Elsevier B.V.; 2014;541: 82–91. doi:10.1016/j.gene.2014.02.060

142. Ding Y, Chang J, Ma Q, Chen L, Liu S, Jin S, et al. Network Analysis of Postharvest Senescence Process in Citrus Fruits Revealed by Transcriptomic and. 2015;168: 357–376. doi:10.1104/pp.114.255711

143. Canas R a., Canales J, Munoz-Hernandez C, Granados JM, Avila C, Garcia-Martin ML, et al. Understanding developmental and adaptive cues in pine through metabolite profiling and co-expression network analysis. J Exp Bot. 2015;66: 3113–3127. doi:10.1093/jxb/erv118

144. Bunyavanich S, Schadt EE, Himes BE, Lasky-Su J, Qiu W, Lazarus R, et al. Integrated Genome-wide Association, Coexpression Network, and Expression Single Nucleotide Polymorphism Analysis Identifies Novel Pathway in Allergic Rhinitis. BMC Med Genomics. 2014;7: 48. doi:10.1186/1755-8794-7-48

145. Matos D a, Cole BJ, Whitney IP, MacKinnon KJ-M, Kay S a, Hazen SP. Daily Changes in Temperature, Not the Circadian Clock, Regulate Growth Rate in Brachypodium distachyon. PLoS One. 2014;9: e100072. doi:10.1371/journal.pone.0100072

146. Wang Z-Y, Tobin EM. Constitutive Expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) Gene Disrupts Circadian Rhythms and Suppresses Its Own Expression. Cell. 1998;93: 1207–1217. doi:10.1016/S0092-8674(00)81464-6

147. Para A, Farré EM, Imaizumi T, Pruneda-Paz JL, Harmon FG, Kay S a. PRR3 Is a vascular regulator of TOC1 stability in the Arabidopsis circadian clock. Plant Cell. 2007;19: 3462–

73. doi:10.1105/tpc.107.054775

148. Handakumbura PP, Matos D a, Osmont KS, Harrington MJ, Heo K, Kafle K, et al. Perturbation of Brachypodium distachyon CELLULOSE SYNTHASE A4 or 7 results in abnormal cell walls. BMC Plant Biol. BMC Plant Biology; 2013;13: 131. doi:10.1186/1471-2229-13-131

149. Vaughn JN, Ellingson SR, Mignone F, Arnim A von. Known and novel post-transcriptional regulatory sequences are conserved across plant families. RNA. 2012;18: 368–84. doi:10.1261/rna.031179.111

150. Kooiker M, Airoldi C a, Losa A, Manzotti PS, Finzi L, Kater MM, et al. BASIC PENTACYSTEINE1, a GA binding protein that induces conformational changes in the regulatory region of the homeotic Arabidopsis gene SEEDSTICK. Plant Cell. 2005;17: 722–729. doi:10.1105/tpc.104.030130

151. Jacob F, Perrin D, Sanchez C, Monod J. The Operon : A Group of Genes Whose Expression is Coordinated by an Operator. Comptes Rendus Des Seances L'Academie Des Sci. 1960;1729: 1727–1729.

152. Martins PK, Ribeiro AP, Cunha BADB da, Kobayashi AK, Molinari HBC. A simple and highly efficient Agrobacterium-mediated transformation protocol for Setaria viridis. Biotechnol Reports. Elsevier B.V.; 2015;6: 41–44. doi:10.1016/j.btre.2015.02.002

153. Liu M, Yang J, Cheng Y, An L. Optimization of soybean (Glycine max (L.) Merrill) in planta ovary transformation using a linear minimal gus gene cassette. J Zhejiang Univ Sci B. 2009;10: 870–876. doi:10.1631/jzus.B0920204

154. Walker JM. Transgenic Plants [Internet]. Life Sciences. 2009. doi:10.1007/978-1-62703-239-1_1

155. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2: 2006.0008. doi:10.1038/msb4100050

156. Giaever G, Nislow C. The Yeast Deletion Collection: A Decade of Functional Genomics. Genetics. 2014;197: 451–465. doi:10.1534/genetics.114.161620

157. Sage RF. The evolution of C 4 photosynthesis. New Phytol. 2004;161: 341–370. doi:10.1046/j.1469-8137.2004.00974.x

158. VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, et al. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature. Nature Publishing Group; 2015; doi:10.1038/nature15714

159. Shulaev V, Sargent D, Crowhurst R, Mockler T, Folkerts O, Delcher A, et al. The genome of woodland strawberry (Fragaria vesca). Nat Genet. 2011;43: 109–116. Available: http://dx.doi.org/10.1038/ng.740

160. Mayer KFX, Rogers J, Dole el J, Pozniak C, Eversole K, Feuillet C, et al. A chromosome-

based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science (80-). 2014;345: 1251788–1251788. doi:10.1126/science.1251788

161. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41: D991–D995. doi:10.1093/nar/gks1193

162. Ransbotyn V, Yeger-Lotem E, Basha O, Acuna T, Verduyn C, Gordon M, et al. A combination of gene expression ranking and co-expression network analysis increases discovery rate in large-scale mutant screens for novel Arabidopsis thaliana abiotic stress genes. Plant Biotechnol J. 2014; 1–13. doi:10.1111/pbi.12274

163. Suwanwela J, Farber CR, Haung B, Song B, Pan C, Lyons KM, et al. Systems genetics analysis of mouse chondrocyte differentiation. J Bone Miner Res. 2011;26: 747–760. doi:10.1002/jbmr.271

164. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, et al. Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight. PLoS Genet. 2006;2: e130. doi:10.1371/journal.pgen.0020130

165. Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S. Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol. 2010;152: 29–43. doi:10.1104/pp.109.145318

166. Hwang S, Rhee SY, Marcotte EM, Lee I. Systematic prediction of gene function in Arabidopsis thaliana using a probabilistic functional gene network. Nat Protoc. Nature Publishing Group; 2011;6: 1429–42. doi:10.1038/nprot.2011.372

167. Schön CC, Utz HF, Groh S, Truberg B, Openshaw S, Melchinger AE. Quantitative Trait Locus Mapping Based on Resampling in a Vast Maize Testcross Experiment and Its Relevance to Quantitative Genetics for Complex Traits. Genetics. 2004;167: 485–498. doi:10.1534/genetics.167.1.485

168. Vogel C, Marcotte EM. Insights into regulation of protein abundance from proteomics and transcriptomis analyses. Nat Rev Genet. 2013;13: 227–232. doi:10.1038/nrg3185.Insights

169. Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri R V, Grotewold E. AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol. 2006;140: 818–29. doi:10.1104/pp.105.072280

170. Steffens NO, Galuschka C, Schindler M, Bülow L, Hehl R. AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. Nucleic Acids Res. 2004;32: D368–72. doi:10.1093/nar/gkh017

171. O'Connor TR, Dyreson C, Wyrick JJ. Athena: A resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. Bioinformatics. 2005;21: 4411–4413. doi:10.1093/bioinformatics/bti714

172. Guo A, He K, Liu D, Bai S, Gu X, Wei L, et al. DATF: A database of Arabidopsis

transcription factors. Bioinformatics. 2005;21: 2568–2569. doi:10.1093/bioinformatics/bti334

173. Barta E, Sebestyen E, Palfy TB, Toth G, Ortutay CP, Patthy L. DoOP: Databases of Orthologous Promoters, collections of clusters of orthologous upstream sequences from chordates and plants. Nucleic Acids Res. 2005;33: D86–90. doi:33/suppl_1/D86 [pii]\r10.1093/nar/gki097

174. Bailey TL, Boden M, Buske F a, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37: W202–8. doi:10.1093/nar/gkp335

175. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Second Int Conf Intell Syst Mol Biol. 1994;2: 28–36. Available: http://www.ncbi.nlm.nih.gov/pubmed/7584402

176. Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Mach Learn. 1995;21: 51–80. doi:10.1007/BF00993379

177. Higo K, Ugawa Y, Iwamoto M, Korenaga T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 1999;27: 297–300. doi:10.1093/nar/27.1.297

178. Chang W-C, Lee T-Y, Huang H-D, Huang H-Y, Pan R-L. PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. BMC Genomics. 2008;9: 561. doi:10.1186/1471-2164-9-561

179. Shahmuradov IA. PlantProm: a database of plant promoter sequences. Nucleic Acids Res. 2003;31: 114–117. doi:10.1093/nar/gkg041

180. Hieno A, Naznin HA, Hyakumachi M, Sakurai T, Tokizawa M, Koyama H, et al. Ppdb: Plant Promoter Database Version 3.0. Nucleic Acids Res. 2014;42: D1188–92. doi:10.1093/nar/gkt1027

181. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, et al. RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Res. 2015;43: W50–W56. doi:10.1093/nar/gkv362

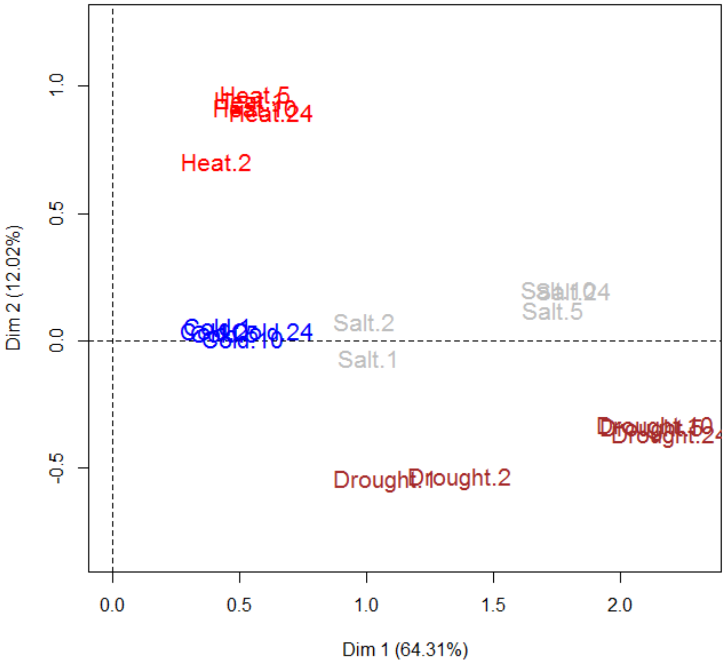
182. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, et al. The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. Nucleic Acids Res. 2008;36: 1009–1014. doi:10.1093/nar/gkm965

183. Matys V. TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003;31: 374–378. doi:10.1093/nar/gkg108

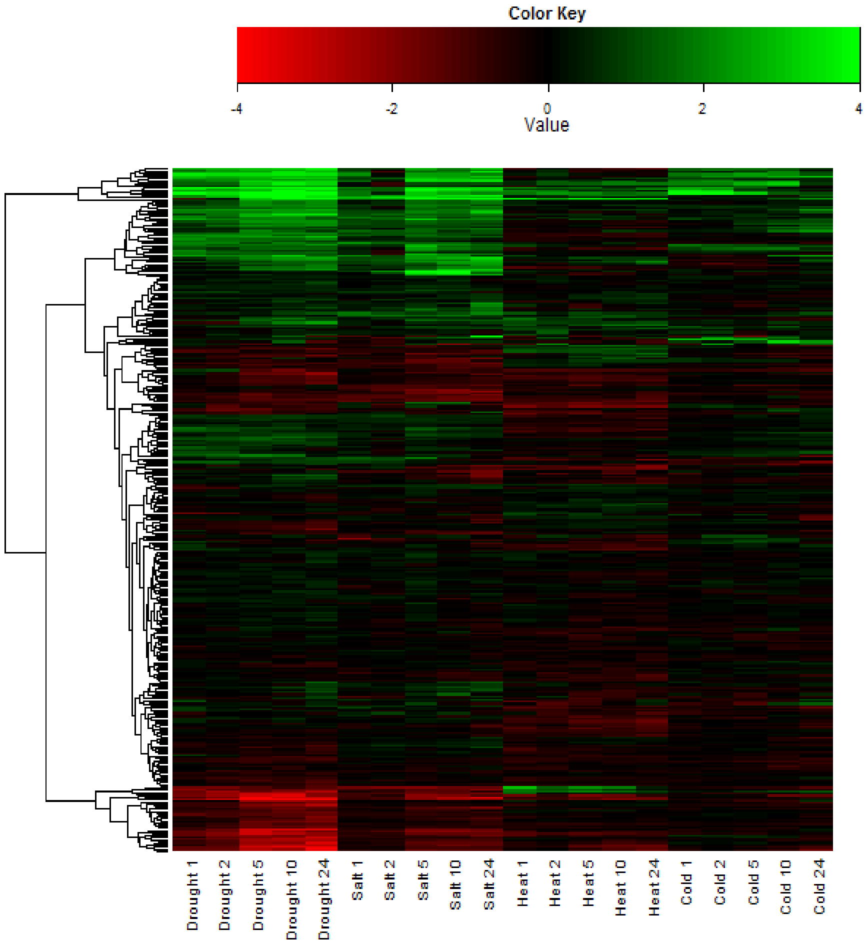
184. Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 2004;32: W199–W203. doi:10.1093/nar/gkh465

Appendix A

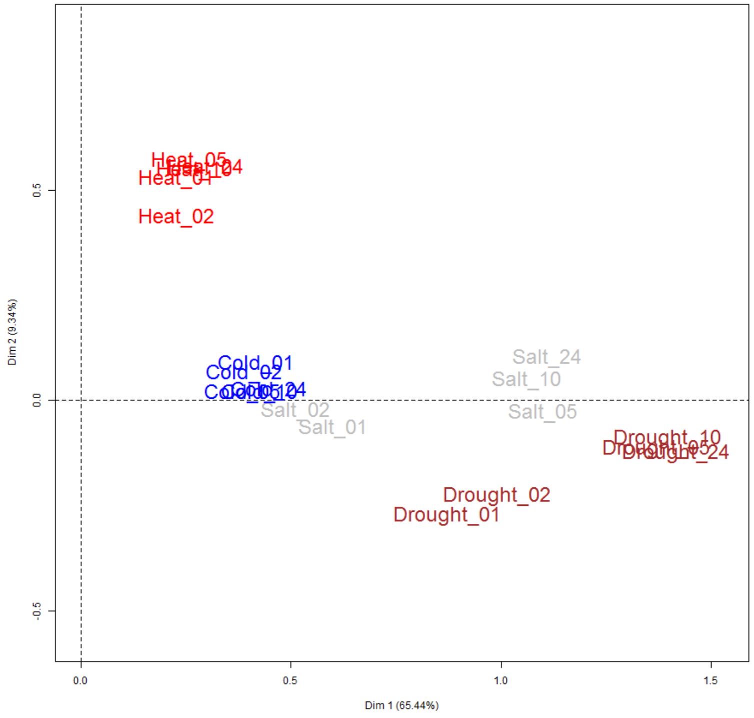
1.1 Supplemental Figures



Supplemental Figure 1. Principal component analysis of RMA normalized microarrays.



Supplemental Figure 2 Heatmap of RMA-expression value differences for 359 calcium ion binding associated loci.



Supplemental Figure 3. Principal component analysis of RMA normalized microarray data for 359 calcium ion binding associated loci.

1.2 Supplemental Files

Supplemental File 1. List of member genes for all co-expressed gene modules in *Brachypodium distachyon* abiotic stress response network

http://www.danforthcenter.org/hpriest/Supplemental_File_1.xlsx

Supplemental File 2. Results of gene ontology functional term enrichment for all gene modules in *Brachypodium distachyon* abiotic stress response network

http://www.danforthcenter.org/hpriest/Supplemental_File_2.xlsx

Supplemental File 3. Results of Element promoter sequence analysis for promoters associated with all gene modules in *Brachypodium distachyon* abiotic stress response network

http://www.danforthcenter.org/hpriest/Supplemental_File_3.xlsx

Supplemental File 4. Software manual and usage guide for dGCNA.

http://www.danforthcenter.org/hpriest/Supplemental_File_4.xlsx

Supplemental File 5. Scale free criteria computed for the comparison of AtLHY-OX and Col-0 datasets

http://www.danforthcenter.org/hpriest/Supplemental_File_5.xlsx

Supplemental File 6. Gene ontology enrichment statistics for High- and Low-connectivity gene groupings in each of the 6 possible networks (total elasticity, positive elasticity, low elasticity, AtLHY-OX, Col-0, and Unified)

http://www.danforthcenter.org/hpriest/Supplemental_File_6.xlsx

Supplemental File 7. Promoter analysis statistics for High- and Low-connectivity gene groupings in each of the 6 possible networks (total elasticity, positive elasticity, low elasticity, AtLHY-OX, Col-0, and Unified)

http://www.danforthcenter.org/hpriest/Supplemental_File_7.xlsx

Supplemental File 8. Gene lists, promoter analysis statistics, and gene ontology enrichment results for all clusters in the positive and negative elasticity networks.

http://www.danforthcenter.org/hpriest/Supplemental_File_8.zip

Supplemental File 9. Gene ontology enrichment statistics for the genes in the immediate neighborhood of AtLHY in the positive and negative elasticity networks.

http://www.danforthcenter.org/hpriest/Supplemental_File_9.xlsx

Supplemental File 10. Criteria determined by permutation and scale free topology for usage in all pairwise comparisons

http://www.danforthcenter.org/hpriest/Supplemental_File_10.xlsx

Supplemental File 11. Positive, negative, and total significant elastic edge numbers for all comparisons

http://www.danforthcenter.org/hpriest/Supplemental_File_11.xlsx