Systems analysis of stress response in plants

Arjun Krishnan

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Genetics, Bioinformatics and Computational Biology

Andy Pereira, Committee Chair

T. M. Murali

Ruth Grene

Allan W. Dickerman

8 September 2010

Blacksburg, Virginia

Keywords: drought response, Arabidopsis, rice, gene expression profiling, regulatory programs, modules, networks, mutant resource, bioenergy

Systems analysis of stress response in plants

Arjun Krishnan

Abstract

The response of plants to environmental stress spans several orders of magnitude in time and space, causing system-wide changes. These changes comprise of both protective responses and adverse reactions in the plant. Stresses like water deficit or drought cause a drastic effect in crop yield, while concomitantly agriculture consumes 1/3rd of the fresh water available to us and there is widespread water scarcity around the world. It is, hence, a fundamental goal of modern biology and applied biotechnology to unravel this complex stress response in laboratory model plants like Arabidopsis and crop models like rice. Such an understanding, especially at the cellular level, will aid in informed engineering of stress tolerance in plants. We have developed and used integrative functional genomics approaches to characterize environmental stress response at various levels of organization including genes, modules and networks in Arabidopsis and rice. We have also applied these methods in problems concerning bioenergy. Since the poor knowledge of the cellular roles of a large portion of plant genes remains a fundamental barrier to using such approaches, we have further explored the problem of 'gene function prediction'. And, finally, as a contribution to the community, we have curated a large mutant resource for the crop model, rice, and established a web resource for exploratory analysis of abiotic stress in this model. All together, this work presents insights into several facets of stress response, offers numerous novel predictions for experimental validation, and provides principled analysis frameworks for systems level analysis of environmental stress response in plants.

Dedication

To my mother

For her uncompromising spirit in giving me the best in life

Acknowledgements

As I see it, this dissertation is a culmination of the efforts of a whole league of people who have nurtured, taught, influenced, encouraged, supported and worked with me over several years.

I'm immensely thankful to my advisor Professor Andy Pereira for making this whole endeavor possible. Thanks Andy for taking me into your hands and patiently supporting me *limitlessly* in these years and giving me plenty of chances to explore, learn and work on diverse problems. I cannot tell you how big an opportunity you have given me. Thanks are also due for creating this great collaborative environment in the lab, without which almost all of this work would not have been possible. I'm grateful to all my current and past lab mates for sharing the good time, especially Utlwang Batlang, Amal Harb, Madan Ambavaram, Peter Wittich and Ankit Gupta with whom I have worked closely. Along with my advisor, while professionally collaborating on projects, they have always respected by unusual perspectives, tolerated my personal idiosyncrasies (e.g. phased-shifted circadian rhythm) and have patiently taught me the little plant biology I know now.

I'm grateful to my committee members Drs. T. M. Murali, Ruth Grene and Allan Dickerman for constant interest and guidance. The collaboration with Murali has been a fantastic experience, and the innumerable discussions I have had with Murali and his students as an 'honorary' guest in his lab meetings have shaped my perspectives in computational systems biology and computer science in general. Dr. Grene has been very kind in helping us interpret our results, discussing them in the light of the big 'plant biology' picture and carefully reading my thesis. Allan has always kept his office doors open for me to drop by and discuss my progress and general ideas.

I'm thankful to Drs. David Bevan, Rahul Kulkarni, Naren Ramakrishnan, Josep Bassaganya-Riera, Eva Collakova, Indra Sandal and Diya Banerjee for the wonderful collaboration we have been engaged in. I have learnt immensely about various other biological systems and thought intently about very diverse problems/questions because of working with these faculty on small projects. External researchers including Drs. Olivier Elemento, Curtis Huttenhower, and

members of the Internaltional Rice Functional Genomics Consortium have been equally kind and helpful to me while working together on exciting projects. I extend my gratitude to the behind-the-scenes developers of open source software tools and databases including R, Bioonductor, MeV, Cytoscape, STAMP, TAIR, Rice Genome Annotation Database, Gramene and Gene Ontology.

I sincerely thank the Genetics, Bioinformatics and Computational Biology program at Virginia Tech for hosting me. Specifically, like everyone in this program, I'm indebted to Dennie Munson, the program coordinator (which means mother and savior for *all*) for holding everything together. My friends Vandana *Chechi*, Yamuna, Kartik, PK and others in Asha for Education have together made my stay in Blacksburg a memorable one and subtracting the time I have spent with them here would make the good time I have had here negligible.

Time before grad school has been as vital as the time at grad school. Throughout my middle and high school days, I was consistently indifferent to the textbooks-assignments-exams-results system and thereby gave a hard time to my parents and class teachers alike. I'm grateful to my teachers in Adarsh Vidhyalaya, and National Public School, Chennai, for protecting me from the system when possible, letting me be on my own when I proved too troublesome, and, instigating my interests in activities outside class like quizzing, drawing, math olympiads and science exhibitions. I'm grateful to everybody at DAV Boys Senior Secondary School, Chennai for showing me an entirely different facet about the *real* world and its ruthless competition, which helped me mature as a person.

I developed my strongest desires about learning – *anything and everything* – during the time I spent with Prof. Venkataraman, my grandfather's friend. Being the epitome of knowledge and not minding the naïveté of a 10^{th} grader, he adopted me as his disciple *and* friend, sharing nuggets of his wisdom with me while teaching me how to solve crossword puzzles, pointing me to books I could read, discussing music and art, and encouraging me to write. It is needless to say how indebted I am to him.

I'm grateful to Profs. Subramanian (TRS), Govindarajan, Santhanam and Ananthan who taught the IIT-JEE coaching classes that I went to during my 11^{th} and 12^{th} grades. While breathing CO_2, sweating to death in a class of 200-odd jam-packed in a normal-sized classroom, and waiting for the eternal words like *'I'm stopping here!'* marking the end of three intense non-stop hours of problem solving, I learnt immensely from these classes and sharpened my skills in mathematics, physics and chemistry. These skills have proven extremely useful to me in my research.

My four years undergraduate studies at the Center for Biotechnology, Anna University, is the one period in life I will never forget for the tremendous influence it had on me on personal and academic fronts. I'm vastly grateful Drs. Geetha Muthukumaran, Goutham Pennathur (G), Karthikeyan Sivaraman (KT), R. B. Narayanan (RBN), Sharmla, Raman sir, Radha ma'am for shaping my interests in science and research. G and RBN graciously opened up their research labs and resources for me and many other students to work in/with them whenever we wished. KT was the graduate student in G's group who was the biggest instigator of my interests in computational biology. Raman sir's depth in knowledge and meticulous teaching influenced me a lot. All of us were very fond of Geetha ma'am and wished (and still wish) to be what she is: an embodiment of knowledge, intelligence, sagacity and lucid communication. The numerous discussions I have had with all these faculty members have made me more and more certain about pursuing science.

My friends have been and are my biggest asset and I gained a number of them during my undergraduate days. I thank all of them, especially Premal, Subramanian (Subu), Prasanna (RSP), Krishnan, Rajagopal and Aneesh, for the tremendously exuberant and joyful times we have had together. Some of the most unforgettable times have been when Premal, Subu and I (occasionally joined by RSP) would spend hours on end talking about everything under the sun and beyond while following a ritualistic nomadic pattern – a super South Indian decoction *kapi* at my place, hours in the sands of Marina beach, dinner at one of our places and a full night out there.

I met my soul mate, Janani Ravi, during my undergrad. Together, we have developed together as individuals, complementing each other all the time, and influencing each other's ways of

thinking. By way of a miracle, we landed in the same university in the same program for PhD! In these eight years, Janani has shown her love in innumerable ways – caring, understanding, generously supporting, pampering, and forgiving. We'll continue on our adventure for the rest of our lives!

I thank my family for giving me plenty of care and support through good and bad times. With a very inconsistent academic performance all the time, I have always kept them guessing about what I could/would become when I grow up. Simply based on their hunch that I could probably do well in the sciences because of my general interests and some math skills they have extended consistent encouragement and help in making things easy for me by providing enough resources all the time. My grand parents (A. R. Parameswaran and T. G. Thangam), aunt's family (Uma Maheswari and R. Muralidharan) and uncle's family (Sivaram and Shobana) have poured unwavering affection on me and have set themselves high up in their careers/lives making me look up to them for inspiration. Rahul and Aditya are loving brothers and their company has always filled me with pleasure.

Janani's parents and family have been equally warm-hearted, and have backed us up all the while (esp. during our GRE preparation), engaged us in constructive tasks (preparing online teaching materials) and provided several advise at the right time that have always made a big impact. Subu's, Rajagopal's, Premals's and RSP's parents have treated me like their son and I'm immensely grateful to all of them.

My father has given me so much of comfort and resources without any questions, and has let me make my own choices all the time, trusting that I would make the right ones. My mother, with so much dedication and uncompromising spirit, has made many a sacrifice to give me the best in life. I'm grateful to my father and mother for bringing me up to what I am right now, and making this meaningful endeavor possible.

Attribution

Several colleagues and coworkers aided in the writing and research behind several of the chapters of this dissertation. A brief description of their background and their contributions are included here.

Prof. Andy Pereira, Ph.D. (Virginia Bioinformatics Institute, Virginia Tech) is the primary Advisor and Committee Chair. Dr. Pereira designed and coordinated all the research presented here. Furthermore, Dr. Pereira contributed in data interpretation and co-wrote the manuscripts.

Chapter 3: Discovering regulatory programs underlying drought response in Arabidopsis.

Amal Harb, Ph.D. (Department of Biology, Virginia Tech) was a student in the author's group and contributed during her graduate studies to this chapter with her drought experiments in Arabidopsis, gene expression profiles, which have been analyzed further. Also, Dr. Harb performed all the validation experiments (indicated, but not included here) and helped in data interpretation.

Chapter 4: The state of network-based gene function prediction in Arabidopsis.

T. M. Murali, Ph.D. (Department of Computer Science, Virginia Tech) worked in collaboration with the author on this project. Dr. Murali, devised the algorithms, implemented the software, ran the cross-validation pipeline and co-wrote the manuscript.

Prof. Brett Tyler, Ph.D. (Virginia Bioinformatics Institute, Virginia Tech) worked with Dr. Murali in devising one of the main algorithms (Sink Source) used in this study.

Chapter 5: A transcriptional regulatory network coordinating activation of cellulose and repression of lignin biosynthesis pathways in rice.

Madana Ambavaram, Ph.D. (Virginia Bioinformatics Institute, Virginia Tech), is a post-doctoral fellow in the author's group and performed all the rice laboratory experiments, helped in data analysis and co-wrote the manuscript.

Kurniawan Trijatmiko, Ph.D., (International Rice Research Institute, Philippines) was a student of Dr. Pereira and contributed during his graduate studies to this chapter with rice transgenic plants.

Chapter 6: Mutant resources in rice for functional genomics of the grasses.

Members from several research groups around the world that are part of the International Rice Functional Genomics Consortium generously shared with us information about their rice mutant collections: **Emmanuel Guiderdoni** (Centre de Cooperation Internationale en Recherche Agronomique pour le Développement, France), **Gynheung An** (Department of Life Science and National Research Laboratory of Plant Functional Genomics, Pohang University of Science and Technology, South Korea), **Yue-ie C. Hsing** (Institute of Plant and Microbial Biology, Academia Sinica, Taiwan), **Chang-deok Han** (Division of Applied Life Sciences, Plant Molecular Biology and Biotechnology Research Center, Gyeongsang National University, South Korea), **Myung Chul Lee** (Rice Functional Genomics, National Institute of Agricultural Biotechnology, South Korea), **Su-May Yu** (Institute of Molecular Biology, Academia Sinica, Taiwan), **Narayana Upadhyaya** (CSIRO Plant Industry, Australia), **Srinivasan Ramachandran** (Temasek Life Sciences Laboratory, National University of Singapore, Singapore), **Qifa Zhang** (National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, China), **Venkatesan Sundaresan** (Plant Biology and Agronomy, University of California), **Hirohiko Hirochika** (Division of Genome and Biodiversity Research, Genetics Department, National Institute of Agrobiological Sciences, Japan), and **Hei Leung** (International Rice Research Institute, Philippines).

Chapter 7: A resource for systems analysis of transcriptional modules involved in drought response in rice.

Madana Ambavaram, Ph.D. (Virginia Bioinformatics Institute, Virginia Tech), is a post-doctoral fellow in the author's group and performed the rice drought experiments and molecular work.

Utlwang Batlang, Ph.D. (Virginia Bioinformatics Institute, Virginia Tech), was a student in the author's group and contributed during his graduate studies to this chapter by working with Dr. Ambavaram in the drought experiments in rice, gene expression profiles, which have been analyzed further.

Table of contents

List of Figures

List of Tables

1. Introduction

Plants, as sessile organisms, have developed a complex repertoire of mechanisms that are unleashed in response to environmental stresses including water deficit or drought. Still, environmental stresses cause large losses in yield and crop performance. The situation is compounded with enormous demands on water resources for agriculture and climatic changes that push towards lesser and lesser available fresh water. One of the primary aspects of any effort to cope with these constraints is a systematic understanding of plant stress responses at various scales of space and time – from single cell to the whole plant level, across growth and developmental stages. Towards this goal, we have pursued research that unravels such responses in Arabidopsis (the model plant) and rice (the crop model) through systems-level profiling of molecular and cellular states during response. The work presented here encapsulates computational systems biology approaches that have been developed and used in conjunction with experimental results to unravel stress response at various levels of organization including genes, modules and networks in Arabidopsis and rice. The following are brief descriptions of the six chapters in this dissertation.

Chapter 1 is a comprehensive survey of the field of integrative functional genomics as applied to understanding molecular mechanisms in Arabidopsis. The review focuses on combining gene expression profiling, regulatory sequence information and functional annotations to mine 'regulatory programs' – groups of cis-regulatory elements (CREs) that potentially mediate the transcriptional regulation of functionally coherent groups of genes.

Chapter 2 describes using the principles presented in the first chapter to discover regulatory programs underlying drought response in Arabidopsis. Gene expression profiles were obtained from plants subjected to a diverse set of drought treatments – progressive, moderate and acute – across time points. 30 known and novel CREs were discovered and combinations of CREs were associated with the drought-regulated expression pattern of specific biological processes.

Chapter 3 deals with 'gene function prediction', a fundamental challenge faced by systems-level efforts in Arabidopsis due to lack of any functional annotations for more than 50% of the genes

in the genome. We have explored the performance of network-based gene function prediction algorithms in predicting 'specific' functions that annotate different numbers of genes annotated to them, have different network topological properties and are differently conserved across diverse species. We have identified several avenues or gaps in functional annotations that can be filled by computational and experimental work in the future.

Chapter 4 takes a detour from stress response to addressing a problem concerning the use of cellulosic plant biomass from rice for bioenergy production. Using a system-level approach combining gene expression profiling and independent global coexpression network analysis, we have discovered a novel role for the transcription factor SHN in coordinately regulating cellulose and lignin biosynthetic pathways. Supported by several experiments at the molecular and phenotypic level, we have arrived at a hypothetical transcriptional regulatory model for lignocellulose biosynthesis in rice.

Chapter 5 presents a rice mutant compendium that contains the mapping of >200,000 insertion sequence tags from 11 international resources to the rice genome, together covering approximately 2/3rd of the protein-coding genes in rice. Furthermore, the coverage of genes of different sizes, and belonging to different biological processes, molecular functions have been surveyed.

Chapter 6 summarizes a community web resource for exploratory functional genomics analysis of abiotic stress in rice. Using drought as an example, based on drought-responsive gene expression profiles across three growth stages and an independent 'environment' coexpression network, we have delineated coexpression modules that are relevant to drought response in the three stages. Moreover, we have cross-referenced each module of genes with functional annotations, regulatory sequence information and genomic positions with respect to abiotic stress QTLs. In addition to revealing the underling module- and network-level response to drought in rice, the resource offers opportunities for prediction of novel genes that could confer drought tolerance.

All supplementary figures and tables are provided at http://treebeard.vbi.vt.edu/AK_Thesis/.

2. Integrative approaches for mining transcriptional regulatory programs in Arabidopsis

Arjun Krishnan, Andy Pereira

Brief Funct Genomic Proteomic. (2008) 7(4): 264-274.

2.1. Abstract

Challenges in modern biology demand a shift in focus from individual gene and protein components to their interacting whole. Integrating information from multiple genomic datasets is seen as a means to this end, capable of providing robust and accurate ways to unravel these functional associations. Integrative strategies, both novel and adapted from other well-studied organisms, are being employed in the model plant *Arabidopsis thaliana* to interpret genome-wide expression, metabolic profiling and protein interaction studies. Exciting inroads are being made in mining and interpretation of developmental, physiological and environmental-response 'programs' using sequence and functional information. The fundamental transcriptional regulatory logic is emerging in Arabidopsis, presently revealed as isolated conditional, spatial or temporal regulatory 'modules'. This immediately calls for efforts towards assembling these building blocks together into a unifying model, thus creating standards for future work to compare with. As a young field, Arabidopsis systems biology is ripe with such an opportunity, now scarcely realizable in other model organisms.

2.2. Introduction

Arabidopsis thaliana, a common weed, has become the primary experimental plant model in a reductionist approach to dissect different aspects of plant biology (Somerville & Koornneef, 2002). Arabidopsis was originally chosen as a genetic model because of its small size, simple growth requirements, short generation time and abundance of seed for genetic analysis. Genetic transformation can now be performed in any laboratory by simply dipping or spraying the young flowers in a bacterial suspension containing a plasmid for transfer to the genome. This methodology has facilitated the generation of millions of insertional mutants that are available

for genetic analysis. Extensive genetic diversity exists in the species in the form of ecotypes adapted to diverse locations around the world, which display tremendous phenotypic variation as well as DNA polymorphisms. As a model for complex eukaryotes with the advantage of being amenable to desired genetic manipulations, and as a reference for crop plants, Arabidopsis unites the understanding of biological phenomena and its application.

The Arabidopsis genome sequence of ~125 Mb size was completed in 2000 (Arabidopsis Genome Initiative, 2000), and has been annotated to contain 27,029 protein coding genes, 3889 pseudogenes/transposable elements and 1123 non-coding RNAs to make a total of 32,041 genes (TAIR7 release, http://www.arabidopsis.org/). Continuing efforts of the Arabidopsis community have resulted in the development of a number of genetic tools such as insertion mutagenesis to mutate all the genes (Alonso et al, 2003), gene expression resources and other functional genomics tools to enable the dissection of a multitude of biological processes (http://www.arabidopsis.org/portals). The objective of the Arabidopsis community is to establish the function of all Arabidopsis genes by the year 2010 (Chory et al, 2000). Meanwhile, the emerging era in plant biology is focused on the genomics-based data collection and individual gene studies to make sense in a systems view.

2.2.1. Too much data! But, not enough data!

Genome-wide high-throughput techniques like DNA sequencing, gene expression, metabolic profiling and large-scale detection of physical interaction have, like in any other organism, revolutionized research and discovery in Arabidopsis. They generate an overwhelming amount of data about the identity and/or levels of various molecules in the cell that make up its cellular state, providing a thorough readout of the system under consideration in high-dimensional space. *But*, each of these methodologies provides a snapshot of a slice through the complex cellular system, only slightly more expansive when these measurements are repeated across a spatial or temporal scale. Also, any of these methods by themselves produce very noisy data (due to experimental and biological variations) that have to be carefully analyzed and interpreted. The golden dream of systems biology – to understand cellular processes as interconnected and interdependent substructures all in the context of biological function and behavior – can be begun to be realized, as a first step, only by drawing from many of these experimental and other

knowledge-based data and integrating them to derive the underlying structure of the system. The basic advantages and themes of data integration are, a) support from multiple distinct evidences increases the confidence in any inference, and b) because different datasets cover different subsets of the whole system, their integration can increase the coverage of the search.

2.2.2. Mining gene expression profiles

Several novel integrative strategies, along with those adapted and extended from other well-studied organisms, are being increasingly used in Arabidopsis research to maximize the interpretation from large-scale experiments, especially genome-wide expression studies. The immense interest in gene expression profiling stems partly from the fact that transcriptional regulation is a favorite point of control, both theoretically and practically. Transcriptional regulation is immediately seen as the conditional turning on or off of genes, which is assumed to directly correlate to presence or absence of their effect inside the cell (see Box 1.1). Not surprising is it, then, that most of the efforts towards engineering cellular systems still focus on directly changing gene expression features (for example (Joshi & Lopez, 2005; Verpoorte & Memelink, 2002)).

Box 1.1: Regulation of gene expression is extremely complex

Regulation of gene expression – the control of the amount and timing of production of the functional product (RNA/protein) – is very intricate in complex organisms, and especially in higher eukaryotes. Gene expression can be modulated in several ways by affecting one of the following cellular processes/states: chemical or structural modification of DNA/chromatin, transcription, RNA transport and degradation, translation, and post-translational modification. So, it is important to understand the immense simplification that is made when gene expression as measured by microarrays is considered to correlate with gene function, when what is assayed is just the regulation of transcription.

The other point to keep in mind is that the process of transcription itself is extremely noisy. See Box 2 of (Komili & Silver, 2008) for an extensive explanation of the sources of noise in transcription.

The final goal of these efforts is to elucidate the transcriptional regulatory network underlying all the processes happening in the cell, covering all the genes in the genome, from regulatory to structural, signaling and catalytic molecules, under all conditions. Results from current studies are lighting up small isolated parts of this network, in the form of *regulatory programs*, each consisting of a few candidate regulatory (transcription factor; TF) genes and their targets tied together using anything from predicted functional associations to verified physical interactions. Many such developmental, physiological and environmental-response programs are being mined in Arabidopsis using integrative approaches that generally involve gene expression correlations, functional classification and regulatory sequence information. Genome-wide gene expression profiling is done mainly using the microarray technology that measures the expression levels of all the queried genes by quantifying the concentrations of the corresponding mRNAs isolated from a sample of a tissue of a given type, in a given condition, at a given time. Functional classification refers to assignment of descriptive terms to genes based on their broad functions or biological processes they participate in. Such classifications have come of age due to concerted efforts to standardize descriptions using controlled vocabulary such as Gene Ontology (GO) (Ashburner et al, 2000). Regulatory sequences here refer to the cis-regulatory elements (CREs) or *motifs* present usually in the proximal promoter region and bound by transcription factors to regulate the expression of adjacent genes (and hence, also called transcription factor binding sites; TFBSs).

In the rest of the review, we will highlight, through different sections, several studies in Arabidopsis that integrate these different datasets using unique and derived approaches to mine regulatory programs. One of our main purposes would be to scan the whole breadth of methodologies applied and the richness of the results obtained.

2.2.3. Coexpression versus Coregulation

The basic assumption of all functional genomic studies on gene expression is that genes with similar expression profiles (coexpression) are regulated by the same mechanisms (coregulation) and participate in the same or similar function. This tenet is supported by quantitative analysis of coexpression data (Allocco et al, 2004) suggesting that both genes with strongly correlated mRNA expression profiles and those with similar functional annotations are more likely to have

their promoter regions bound by a common TF. It is also found that combining expression data with functional annotation results in a better predictive model than using either data source alone.

Coexpression between two genes (correlation between their expression profiles) is quantified using a distance metric measuring the distance between their expression vectors, which for each gene is an array of values of expression of that gene across many experiments. Such a measure can be used to build coexpression networks of genes where genes are connected to each other when they have a high degree of correlation across experiments. Such networks can be readily used to generate several hypotheses about other members of the same pathway, physical interaction partners, regulators etc. Coexpression networks are being used extensively in Arabidopsis research, for assigning gene functions to undefined genes (Rautengarten et al, 2005) or novel pathway members (Persson et al, 2005). Using coexpression data for knowledge discovery in general has been recently reviewed (Aoki et al, 2007), and therefore only specific applications are discussed below.

2.3. Inferring regulatory programs

Microarray experiments produce multi-dimensional expression readouts of all the genes in a genome (those accessed by the platform) and it is hard to immediately discern the meaningful set of genes and hence the biological processes and pathways coordinating the cellular responses recorded by the experiment. From the raw data, expression levels of individual genes have to be obtained after background correction, normalization, and summarization procedures. For the comparison of interest between the two biological states (e.g. treatment vs. control; mutant vs. wild-type) an expression ratio of each gene is then obtained and tested for statistical significance to declare the consistently changed genes (across replications, based on their *p*-value at a desired level of significance after correcting for multiple hypothesis testing) as differentially expressed. So, for every comparison, all genes have an expression ratio associated with it and a subset of the genes is declared ‘differentially expressed’. Two excellent reviews (Clarke & Zhu, 2006; Nettleton, 2006) written for plant biologists discuss these ideas and their practical considerations more rigorously.

2.3.1. Gene expression clustering coupled with whole-genome functional annotations

The lists of differentially up- and down-regulated genes are usually interpreted by searching for the presence and distribution of genes in different functional categories (e.g. GO or gene families) (Pina et al, 2005). The significance and rank of the association of any category with the regulated genes is assessed using a test for enrichment like the one-tailed Fisher's exact test (Fisher, 1922), commonly implemented based on the cumulative hypergeometric distribution. Such an analysis allows the decomposition of the up- and down-regulated lists into clusters of genes that participate in a similar biological process or pathway (see Box 2.2). Several web-based tools are available for performing GO enrichment analysis in Arabidopsis including CLENCH (Shah & Fedoroff, 2004), EasyGO (Zhou & Su, 2007), and FatiGO (Al-Shahrour et al, 2004).

Box 2.2: Gene set enrichment analysis

To interpret a given comparison of gene expression, the conventional approach is to delineate and focus on a list of differentially expressed genes based on a p-value cutoff after statistical testing and correction for testing multiple hypotheses. There are a few major limitations in this approach:

1. After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are modest relative to the noise inherent to the microarray technology.
2. Alternatively, one may be left with a long list of statistically significant genes without any unifying biological theme, in which case interpretation can be daunting and ad hoc.
3. Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting in concert. An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.
4. When different groups study the same biological system, the list of statistically significant genes from the two studies may show distressingly little overlap.

To overcome these limitations a gene set enrichment analysis (GSEA) was proposed (Subramanian et al, 2005), which provides an enrichment score that quantifies the degree of enrichment of a gene set at the top or bottom of an ordered gene list derived from the

experimental data set. The ordered list is the list of differentially expressed genes ranked according to their level of significance (increasing *p*-values), and gene sets are defined based on prior biological knowledge (for example, GO categories, members of a metabolic pathway, targets of a given TF etc.). Other statistical methods of performing GSEA are continually being developed (Backes et al, 2007b; Kim & Volsky, 2005). So, for example, instead of selecting for differentially expressed genes using a *p*-value cutoff, GO enrichment could be carried out using the total set of rank-ordered genes (Zhang et al, 2008), as is being increasingly used in Arabidopsis gene expression research (Diet et al, 2006). Tools like GeneTrail (Backes et al, 2007a) and FatiScan (Al-Shahrour et al, 2006) provide GSEA capabilities for Arabidopsis data.

A more powerful approach is to cluster the differentially expressed genes based on their gene expression profile similarity, and test these clusters for functional enrichment. This was demonstrated by a comprehensive analysis of a large dataset of microarray expression profiles (Schmid et al, 2005) using clustering, principal component analysis (PCA), GO category and gene family enrichment, and expression correlation to gain insights into the regulatory programs controlling development in Arabidopsis. The use of clustering originates from the assumption that genes with similar expression profiles have some characteristic in common, such as being functionally related (e.g. involved in a particular biological process or pathway), or being regulated by the same set of transcription factors. Clustering also makes the data more comprehensible and interpretable. Several clustering algorithms have been proposed and the choice depends on the data (Datta & Datta, 2006; Thalamuthu et al, 2006). A combination of these methods has also been suggested to be useful (Wang et al, 2004). Hierarchical clustering and PCA, which model linear relationships between genes, have been extensively used to analyze Arabidopsis data, but non-linear dimension reduction techniques are being introduced that are suggested to be superior to these conventional approaches (Katagiri & Glazebrook, 2003).

Many tools offer capabilities for clustering gene expression in general, but only a handful among them is compatible with Arabidopsis data when additional analysis of the resultant clusters is required. This is primarily due to lack of Arabidopsis gene annotations within the system. The TM4 suite (Saeed et al, 2006) was one of the first microarray data analysis toolboxes packed

with several tools for classification, dimension reduction, functional annotation and statistical analysis used in Arabidopsis gene expression analysis. Software like MAPMAN (Thimm et al, 2004; Usadel et al, 2005) are extensively used by Arabidopsis researchers to project their expression data onto pathway structure (see below), and this tool also allows clustering of the expression data and using each derived cluster for pathway enrichment analysis and visualization. Several of the high quality microarray datasets in Arabidopsis are from time-course experiments (e.g. AtGenExpress (Kilian et al, 2007)), where gene expression changes in response to a particular biological process or condition has been recorded for a few (three to eight) time points. In such short time-course experiments, many genes may have the same expression pattern just by random chance, and there are no time series repeats for cycling processes. Software like STEM (Ernst & Bar-Joseph, 2006) are well-suited for such data, while they also take into account the temporal nature of the experiment unlike general clustering methods. STEM, in particular, also integrates GO enrichment analysis. As previously noted, for general grouping of gene expression data, hierarchical agglomerative clustering is widely used, but the result is one continuous nested tree connecting all genes. Distinct clusters are usually discerned by cutting the tree at a particular level, or at different levels based on the degree of similarity between levels. The software MultiGO (Kankainen et al, 2006) offers another dimension to such analysis by using GO annotations to identify an optimal cut in the tree that maximizes functional enrichment of all the severed sub-trees.

Performing hierarchical clustering, for instance, on both the genes and samples will allow the identification of global expression signatures: genes that show correlated expression across experimental conditions related by some identifiable criterion such as tissue type, developmental/differentiation state, or signaling response. But, since the clustering of genes (or conditions) is based on the correlation across all conditions (or genes), subsets of coregulated genes that might only be coexpressed in a subset of the conditions (displaying almost complete independence in the rest) will be lost. Classical clustering algorithms also have the drawback of assigning each gene to a single cluster, while many genes can be involved in different biological processes depending on the cellular context and, therefore, they might be coexpressed with different groups of genes under different conditions. Solutions to such challenges have been proposed and applied to Arabidopsis datasets. Fuzzy k-means clustering (Gasch & Eisen, 2002),

capable of identifying overlapping clusters, has been used to analyze an integrated compendium of stress gene expression datasets to elucidate clusters of genes characterizing the response to individual and combination of all stresses (Ma & Bohnert, 2007). The use of biclustering methods, which perform clustering in the two gene-condition dimensions simultaneously, has also been initiated in Arabidopsis to mine local expression signatures (Bleuler et al, 2004; Prelic et al, 2006).

Many time-course and conditional response experiments exist for other species, which are hard to compare because of discrepancies in the platform or non-identical experimental conditions and controls. A unique resource in Arabidopsis – AtGenExpress (Kilian et al, 2007) – contains genome-wide expression profiling data from a large and diverse collection of conditional time-course experiments (encompassing a wide range of abiotic and biotic treatments, the application of plant hormones and chemical treatments) using the same technology platform and reference conditions. This data can immediately be seen as three-dimensional gene-condition-time profiles. This 3D dataset has been analyzed (Strauch M, 2007) using a two-step clustering approach, to first cluster the genes over all time-points under one condition and then, as a second step, follow the identified significant clusters through the condition dimension. In this manner, the initially discovered clusters were classified into coherent (similar response under all conditions), single-response (specific response to a single condition), or individual-response (coregulated under different conditions with different response patterns) programs, depending on the profile patterns that emerge when viewed across conditions. Biological processes captured by these programs were identified using GO categories.

2.3.2. Promoter analysis and assignment of regulatory mechanism

Evolutionarily increasing diversity of TFBSs/CREs is thought to cause the progressively complex gene expression correlating with organismal complexity (Levine & Tjian, 2003), and these elements are believed to encode in the genome, the organisms response diversity – the spectrum of gene expression states induced by different environmental and extracellular conditions (Harbison et al, 2004). Genes that are differentially expressed in response to a large number of different external stimuli are therefore expected to contain more distinct CREs in their upstream regions than are genes that respond to only few environmental cues. This has been

confirmed in Arabidopsis (Walther et al, 2007) where multi-stimuli response genes have been shown to contain an increased absolute CRE count, CRE density and have more paralogs, which might be important for "shifted and novel response scope" of these genes in addition to allowing a greater dynamic expression range.

Hence, characterization of the promoter sequences of the coexpressed genes to delineate their CRE compositions would give additional insights into the exact regulatory mechanism that drives the expression of these genes. If these CREs are associated with particular TFs or TF families, then their composition can be directly used to propose candidate regulatory molecules (Meier et al, 2008). Most microarray studies take this additional step of assigning a combination of CREs to infer cis-regulatory programs or 'modules' from functionally enriched gene expression patterns (Hannah et al, 2005) (Fig. 2.1).

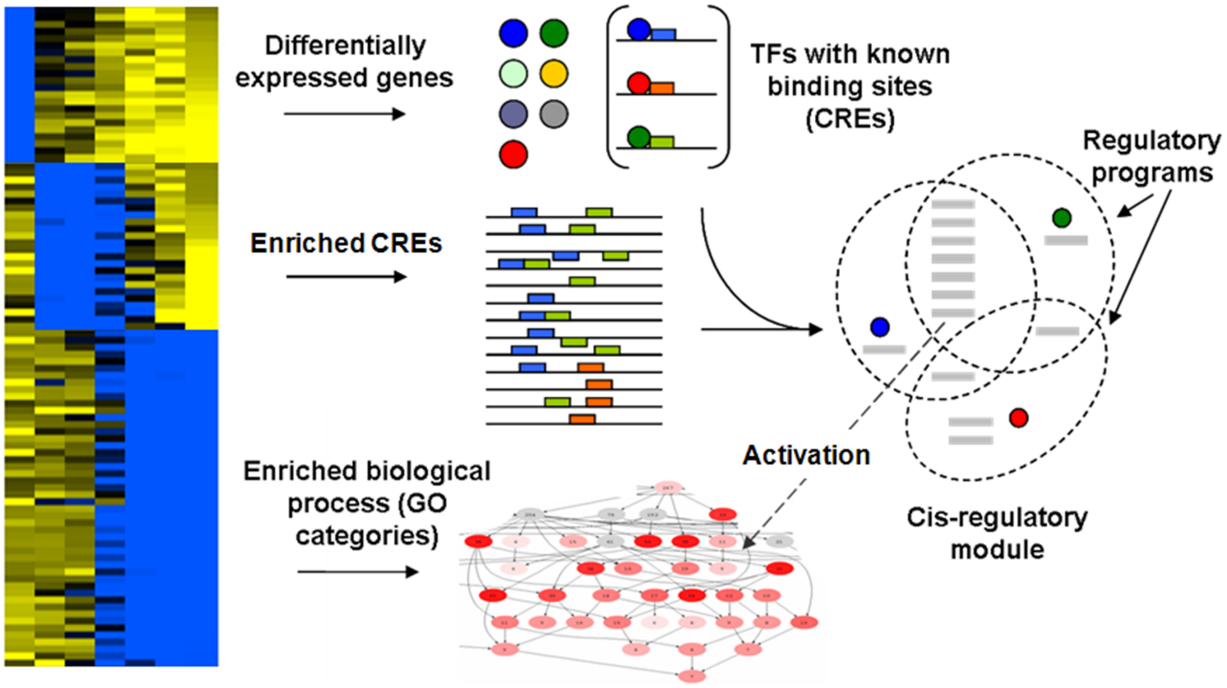


Figure 2.1: An interpretation of a regulatory program, presented along with the possible steps involved in mining it from gene expression profiles.

Relevant CREs are usually identified by starting with experimentally verified elements and using them to seed the search to identify subsets of them that are enriched among coexpressed genes. Novel CREs are similarly identified among coexpressed genes using enrichment of 3-8mer sequences. Both known and novel CREs can be verified for functional significance by checking

for their conservation across species, the assumption being that orthologous genes exhibit a common regulatory mode (Haberer et al, 2006).

Computational algorithms, including machine learning approaches (Li et al, 2006b), have been proposed for Arabidopsis that use coexpression and CRE composition data to identify motifs that provide discriminatory information for predicting the expression patterns of genes. Two other complementary methods to mine such relevant CREs have been proposed (Benedict et al, 2006; Geisler et al, 2006). In the first approach, coregulated gene lists are searched for enriched CREs. For each candidate CRE, all the genes in the genome containing that element are identified and tested against expression data to confirm or reject that the element is correlated with induction or repression of gene expression. In the second method, the set of genes in the genome containing a particular CRE is first recovered, noting for its behavior in different microarray experiments. CREs are deemed 'functional' when their set of genes contain a number of induced/repressed genes significantly higher than from a similarly sized set chosen randomly from the genome.

A biclustering approach has been used to mine cis-regulatory modules in the 3D AtGenExpress dataset (Supper et al, 2007). The three program categories defined in the previous work (Strauch M, 2007) were used as definitions for a random sampling of the dataset. Each such sample was iteratively refined (by removing gene and conditions) and stored if it matches a program definition. Recovered clusters were then merged (to remove redundancy) and extended to ensure gene and condition maximality, and then associated with GO categories and CREs based on their enrichment among the constituent genes. In another study, the generic stress response in Arabidopsis was elucidated on a global scale by integrating and analyzing gene expression profiles measured under different abiotic, biotic and chemical stresses (Ma & Bohnert, 2007). Fuzzy k-means clustering was used to resolve the genes into 180 coexpression groups, which were then subjected to functional (GO) and CRE enrichment analysis. Several known and novel pathways and motifs emerged that together define the ubiquitous and specific stress-response modules. The stress-response modules were finally used to characterize root-specific response to stress by intersecting them with clusters similarly resolved from root gene expression data. A simple and elegant methodology was applied to infer cold acclimation modules by logically superimposing CRE composition and (experimentally determined) TF-CRE correspondence onto

time-course gene expression profiles (Chawade et al, 2007). First, the CREs (associated with a TF) overrepresented among the differentially expressed were identified and the genes containing a common set of CREs were deemed putative target clusters. Targets in each such cluster that were (significantly) regulated in the same time point or in the time-point following their TF were then chosen for evaluating their expression coherence. Subsets of these highly coexpressing genes were finally declared as being regulated by the TF.

Associating multiple, rather than individual CREs to each gene in a coexpression group lends more support to its correlated expression arising from a common regulation program. Such CRE-gene combinations have been modeled as regulatory bicliques, mined using biclustering of a CRE-gene matrix and assessed for significance based on a combination of the multi-CRE enrichment score and the expression coherence of the target genes (Pati et al, 2006). The tools *ModuleFinder* and *CoReg* were developed to identify gene-condition biclusters (see above) with a hierarchically clustered gene tree that is then used to assign a CRE at different levels of the tree (Holt et al, 2006). This is done assuming that the tree structure of the expression data is a reflection of patterns of CREs in promoter regions of the gene involved. In another effort to identify generic regulatory modules in Arabidopsis, a two-way clustering procedure (Vandepoele et al, 2006) was employed to build on a previous strategy . Starting from a set of 34 CREs enriched in sets of genes coexpressed across several microarrays, clusters of genes with similar CRE combinations in their promoters were delineated. Next, within each such cluster, groups of coexpressed genes were identified. Finally, motif detection was applied and CREs evolutionarily conserved (across Arabidopsis and a related species, poplar) were retained. In a very recent study (Michael et al, 2008), a seamless pipeline involving expression pattern discovery and promoter analysis was used to characterize CREs involved in the cyclic regulation of gene expression by light and circadian clock.

Several web-servers house CRE information for Arabidopsis genes (Higo et al, 1999) and provide tools to identify and visualize CREs enriched in a set of genes of interest to the user (Davuluri et al, 2003; Galuschka et al, 2007; O'Connor et al, 2005; Pavesi et al, 2006). Other web-tools combine either clustering of gene expression data or identification of coexpressed

genes to analysis of CRE content (Lescot et al, 2002; Manfield et al, 2006; Obayashi et al, 2009; Toufighi et al, 2005).

Box 2.3: Meta-analysis

Where the term "analysis" is used to describe the quantitative approaches to draw useful information from raw data, the term "meta-analysis" refers to the approaches used to draw useful information from the results of previous analyses (Stevens & Doerge, 2005). Meta-analysis of gene expression involves surveying microarray data across experiments to find the true effect of a treatment across different studies. One of the benefits of a meta-analysis is, hence, also one of the benefits of pooling raw data – increased power to detect significant differences. Combining estimates of differential expression has been found to be more accurate than estimating from individual experiments (Choi et al, 2003).

Meta-analyses have been used to identify "Integration-Driven Revisions" (IDRs), genes identified as differentially expressed by multiple studies or labs, but determined by the results of a meta-analysis to be not differentially expressed. Such genes might be advanced by multiple groups for further study due to their large and significant 'effect size' (Choi et al, 2003), while the meta-analysis concludes that, due to the inconsistencies in estimates across studies, the gene is not significantly differentially expressed. IDRs will, hence, tend to occur when there are *large but inconsistent* estimates. "Integration-Driven Discoveries" (IDDs) have also been recovered when *small but consistent* effect size estimates were combined across studies to identify significant genes that were missed by individual studies.

As the gene expression resources for Arabidopsis keep growing enormously, meta-analysis would greatly aid in refining the findings from a new experiment. The Meta-Analyzer utility in the tool Genevestigator (Zimmermann et al, 2004) makes it possible to study the gene expression profiles of several genes simultaneously in the context of those in environmental stresses, organs, and growth stages. In addition to just using known expression profiles, characterizing the overlap of the genes differentially expressed in previous experiments with those identified in a new gene expression comparison has been suggested (Nielsen et al, 2007). Similar approaches have been previously used to identify, among other things, known expression signatures significantly overlapping with the newly identified ones (Rhodes et al, 2007).

2.4. Conclusion

All methods discussed above give rise to regulatory programs in Arabidopsis to varying degrees of detail. While some simply classify genes into coexpression clusters participating in different biological processes, others have been able to assign candidate CREs to sketch out regulatory modules. Still others have gone further to assign a TF (or a TF family) to these modules based on temporal evidences. All the regulatory programs identified thus far have aided in understanding of cellular response in Arabidopsis to external stimuli and internal physiological changes. It is easy to imagine that each of these modules is a highlighted portion of the underlying global transcriptional regulatory network active in a specific condition, time or cell-type. To foster systematic growth of systems-level knowledge in Arabidopsis, these modules have to be unified, annotated and deposited as a single resource amenable for meta-analyses (see Box 2.3) to serve as a reference for any future integrative bioinformatics work. This possibility of being able organize all the discovered knowledge under one roof exists in the case of Arabidopsis as against other model organisms where a huge number of methods and results have already been produced, making it almost impossible to reconcile all the data.

The global regulatory network can then be constructed by sequentially integrating experimentally verified (de Folter et al, 2005; Gong et al, 2007) and predicted (Geisler-Lee et al, 2007; Yu et al, 2004) large-scale protein-DNA and protein-protein interactions (physical interaction network) along with other 'influence' interactions between genes (the gene-gene network). Such integration can be undertaken using frameworks (Hasegawa et al, 2006; Kohler et al, 2006; Li et al, 2006a) for systematic consolidation of diverse data-types. Recent efforts in this direction have helped shed light on transcriptional (Palaniswamy et al, 2006) and protein interactions (Cui et al, 2008) on a genome-wide scale. Data relating to transcriptional regulation, promoter sequences of the genes and their cis-elements, and transcription factors and their targets, has been brought under one roof (Palaniswamy et al, 2006). From an example in humans (Rhodes et al, 2005), a similar probabilistic approach has been taken to bring together various evidences to predict protein interactions (Cui et al, 2008). Pathway information is also being made increasingly available for Arabidopsis in integrated databases (Tsesmetzis et al, 2008). Analysis methods that take advantage of the connectivity of the genes in these networks and

pathways (Draghici et al, 2007; Khatri et al, 2008) (over treating them merely as sets of genes; see Box 2.2) would help in mining more knowledge about how genes function together to carry out biological processes.

2.4.1. Synthesis towards understanding and application

Unifying data and concepts from model plants to apply to the diversity of crop plants is also taking place at a rapid rate with the availability of draft genome sequences and subsequent functional genomics tools. The role of Arabidopsis is to provide a reference genome with functions of genes established by experimentation that can be extrapolated using comparative genomics to diverse plants.

It is generally accepted that many plant traits such as yield and resistance to biotic/abiotic stresses are complex and can not be only understood at the single gene level. Therefore, the expectation is that Arabidopsis systems biology can provide working models to understand and apply knowledge gained for plant improvement.

2.5. References

Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20:** 578-580

Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, Conde L, Blaschke C, Vera J, Dopazo J (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Research* **34:** W472-W476

Allocco DJ, Kohane IS, Butte AJ (2004) Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **5:** 18-18

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R (2003) Genome-Wide Insertional Mutagenesis of Arabidopsis thaliana. *Science* **301:** 653-657

Aoki K, Ogata Y, Shibata D (2007) Approaches for Extracting Practical Information from Gene Co-expression Networks in Plant Biology. *Plant Cell Physiol* **48:** 381-390

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408:** 796-815

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

Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E, Lenhof HP (2007a) GeneTrail-advanced gene set enrichment analysis. *Nucleic Acids Research* **35:** W186-W192

Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E, Lenhof HP (2007b) GeneTrail–advanced gene set enrichment analysis/. *Nucleic Acids Research* **35:** W186–W192

Benedict C, Geisler M, Trygg J, Huner N, Hurry V (2006) Consensus by democracy. Using meta-analyses of microarray and genomic data to model the cold acclimation signaling pathway in Arabidopsis. *Plant Physiology* **141:** 1219-1232

Bleuler S, Prelic A, Zitzler E (2004) An EA framework for biclustering of gene expression data. *Evolutionary Computation, 2004 CEC2004 Congress on* **1:** 166-173

Chawade A, Bräutigam M, Lindlöf A, Olsson O, Olsson B (2007) Putative cold acclimation pathways in Arabidopsis thaliana identified by a combined analysis of mRNA co-expression patterns, promoter motifs and transcription factors. *BMC Genomics* **8:** 304

Choi JK, Yu U, Kim S, Yoo OJ (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19:** 84-90

Chory J, Ecker JR, Briggs S, Caboche M, Coruzzi GM, Cook D, Dangl J, Grant S, Guerinot ML, Henikoff S (2000) National Science Foundation-Sponsored Workshop Report:" The 2010 Project" functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiology* **123:** 423-426

Clarke JD, Zhu T (2006) Microarray analysis of the transcriptome as a stepping stone towards understanding biological systems: practical considerations and perspectives. *The Plant Journal* **45:** 630-650

Cui J, Li P, Li G, Xu F, Zhao C, Li Y, Yang Z, Wang G, Yu Q, Li Y, Shi T (2008) AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology. *Nucleic Acids Research* **36:** D999-1008

Datta S, Datta S (2006) Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics* **7:** S17

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* **4:** 25

de Folter S, Immink RG, Kieffer M, Parenicova L, Henz SR, Weigel D, Busscher M, Kooiker M, Colombo L, Kater MM (2005) Comprehensive interaction map of the Arabidopsis MADS box transcription factors. *The Plant cell* **17:** 1424-1433

Diet A, Link B, Seifert GJ, Schellenberg B, Wagner U, Pauly M, Reiter WD, Ringli C (2006) The Arabidopsis root hair cell wall formation mutant Irx 1 is suppressed by mutations in the RHM 1 gene encoding a UDP-L-rhamnose synthase. *The Plant cell* **18:** 1630-1641

Draghici S, Khatri P, Tarca A, Amin K, Done A, Voichita C, Georgescu C, Romero R (2007) A systems biology approach for pathway level analysis. *Genome Research* **17:** 1537-1545

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

Fisher RA (1922) On the interpretation of c2 from contingency tables, and on the calculation of P. *J R Stat Soc* **85:** 81-94

Galuschka C, Schindler M, Bulow L, Hehl R (2007) AthaMap web tools for the analysis and identification of co-regulated genes. *Nucleic Acids Research* **35:** D857-D862

Gasch AP, Eisen MB (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology* **3:** 0059.0051-0059.0022

Geisler M, Kleczkowski LA, Karpinski S (2006) A universal algorithm for genome-wide in silico identification of biologically significant gene promoter putative cis-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in Arabidopsis. *The Plant Journal* **45:** 384-398

Geisler-Lee J, O'Toole N, Ammar R, Provart NJ, Millar AH, Geisler M (2007) A predicted interactome for Arabidopsis. *Plant Physiology* **145:** 317-329

Gong W, He K, Covington M, Dinesh-Kumar SP, Snyder M, Harmer SL, Zhu YX, Deng XW (2007) The development of protein microarrays and their applications in DNA protein and protein protein interaction analyses of Arabidopsis transcription factors. *Molecular Plant* **1:** 27-41

Haberer G, Mader MT, Kosarev P, Spannagl M, Yang L, Mayer KFX (2006) Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea. *Plant Physiology* **142:** 1589-1602

Hannah MA, Heyer AG, Hincha DK (2005) A global survey of gene regulation during cold acclimation in Arabidopsis thaliana. *PLoS Genet* **1:** e26

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

Hasegawa Y, Seki M, Mochizuki Y, Heida N, Hirosawa K, Okamoto N, Sakurai T, Satou M, Akiyama K, Iida K, Lee K, Kanaya S, Demura T, Shinozaki K, Konagaya A, Toyoda T (2006) A flexible representation of omic knowledge for thorough analysis of microarray data. *Plant Methods* **2:** 5-5

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

Holt KE, Millar AH, Whelan J (2006) ModuleFinder and CoReg: alternative tools for linking gene expression modules with promoter sequences motifs to uncover gene regulation mechanisms in plants. *Plant Methods* **2:** 8-8

Joshi L, Lopez LC (2005) Bioprospecting in plants for engineered proteins. *Current opinion in plant biology* **8:** 223-226

Kankainen M, Brader G, Toronen P, Palva ET, Holm L (2006) Identifying functional gene sets from hierarchically clustered expression data: map of abiotic stress regulated genes in Arabidopsis thaliana. *Nucleic Acids Research* **34:** e124-e124

Katagiri F, Glazebrook J (2003) Local Context Finder (LCF) reveals multidimensional relationships among mRNA expression profiles of Arabidopsis responding to pathogen infection. *Proceedings of the National Academy of Sciences* **100:** 10842-10847

Khatri P, Draghici S, Tarca A, Hassan S, Romero R (2008) A System Biology Approach for the Steady-State Analysis of Gene Signaling Networks. In *Progress in Pattern Recognition, Image Analysis and Applications*, pp 32-41.

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

Kim SY, Volsky DJ (2005) PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics* **6:** 144

Kohler J, Baumbach J, Taubert J, Specht M, Skusa A, Ruegg A, Rawlings C, Verrier P, Philippi S (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* **22:** 1383-1390

Komili S, Silver PA (2008) Coupling and coordination in gene expression processes: a systems biology view. *Nat Rev Genet* **9:** 38-48

Lescot M, DÇhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, RouzÇ P, Rombauts S (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Research* **30:** 325-327

Levine M, Tjian R (2003) Transcription regulation and animal diversity. *Nature* **424:** 147-151

Li J, Li X, Su H, Chen H, Galbraith DW (2006a) A framework of integrating gene relations from heterogeneous data sources: an experiment on Arabidopsis thaliana. *Bioinformatics* **22:** 2037-2043

Li Y, Lee KK, Walsh S, Smith C, Hadingham S, Sorefan K, Cawley G, Bevan MW (2006b) Establishing glucose-and ABA-regulated transcription networks in Arabidopsis by microarray

analysis and promoter classification using a Relevance Vector Machine. *Genome Research* **16:** 414-427

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

Manfield IW, Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR (2006) Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic Acids Research* **34:** W504-W509

Meier S, Bastian R, Donaldson L, Murray S, Bajic V, Gehring C (2008) Co-expression and promoter content analyses assign a role in biotic and abiotic stress responses to plant natriuretic peptides. *BMC Plant Biology* **8:** 24

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

Nettleton D (2006) A discussion of statistical methods for design and analysis of microarray experiments for plant scientists. *The Plant cell* **18:** 2112-2121

Nielsen HB, Mundy J, Willenbrock H (2007) Functional associations by response overlap (FARO), a functional genomics approach matching gene expression phenotypes. *PLoS ONE* **2:** e676

O'Connor TR, Dyreson C, Wyrick JJ (2005) Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics* **21:** 4411-4413

Obayashi T, Hayashi S, Saeki M, Ohta H, Kinoshita K (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Research* **37:** D987-991

Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E (2006) AGRIS and AtRegNet. A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks. *Plant Physiol* **140:** 818-829

Pati A, Vasquez-Robinet C, Heath LS, Grene R, Murali TM (2006) XcisClique: analysis of regulatory bicliques. *BMC Bioinformatics* **7:** 218-218

Pavesi G, Mereghetti P, Zambelli F, Stefani M, Mauri G, Pesole G (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Research* **34:** W566-W570

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences* **102:** 8633-8638

Pina C, Pinto F, Feijo JA, Becker JD (2005) Gene family analysis of the Arabidopsis pollen transcriptome reveals biological implications for cell growth, division control, and gene expression regulation. *Plant Physiology* **138:** 744-756

Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22:** 1122-1129

Rautengarten C, Steinhauser D, Bussis D, Stintzi A, Schaller A, Kopka J, Altmann T (2005) Inferring hypotheses on functional relationships of genes: analysis of the Arabidopsis thaliana subtilase gene family. *PLoS Computational Biology* **1:** 297-312

Rhodes DR, Kalyana-Sundaram S, Tomlins SA, Mahavisno V, Kasper N, Varambally R, Barrette TR, Ghosh D, Varambally S, Chinnaiyan AM (2007) Molecular concepts analysis links tumors, pathways, mechanisms, and drugs. *Neoplasia* **9:** 443-454

Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM (2005) Probabilistic model of the human protein-protein interaction network. *Nat Biotech* **23:** 951-959

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J (2006) TM4 microarray software suite. *Methods Enzymol* **411:** 134-193

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

Shah NH, Fedoroff NV (2004) CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics* **20:** 1196-1197

Somerville C, Koornneef M (2002) A fortunate choice: the history of Arabidopsis as a model plant. *Nat Rev Genet* **3:** 883-889

Stevens JR, Doerge RW (2005) Combining Affymetrix microarray results. *BMC Bioinformatics* **6:** 57

Strauch M SJ (2007) A two-step clustering for 3-D gene expression data reveals the main features of the Arabidopsis stress response. *Journal of Integrative Bioinformatics* **4:** 54-54

Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub T, Lander E, Mesirov J (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 15545-15550

Supper J, Strauch M, Wanke D, Harter K, Zell A (2007) EDISA: extracting biclusters from multiple time-series of gene expression profiles. *BMC Bioinformatics* **8:** 334

Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* **22:** 2405-2412

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal* **37:** 914-939

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-Northerns, expression angling, and promoter analyses. *The Plant Journal* **43:** 153-163

Tsesmetzis N, Couchman M, Higgins J, Smith A, Doonan JH, Seifert GJ, Schmidt EE, Vastrik I, Birney E, Wu G, D'Eustachio P, Stein LD, Morris RJ, Bevan MW, Walsh SV (2008) Arabidopsis reactome: a foundation knowledgebase for plant systems biology. *Plant Cell* **20:** 1426-1436

Usadel B, Nagel A, Thimm O, Redestig H, Blaesing OE, Palacios-Rojas N, Selbig J, Hannemann J, Piques MC, Steinhauser D (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of coresponding genes, and comparison with known responses. *Plant Physiology* **138:** 1195-1204

Vandepoele K, Casneuf T, Van de Peer Y (2006) Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biology* **7:** R103

Verpoorte R, Memelink J (2002) Engineering secondary metabolite production in plants. *Current opinion in biotechnology* **13:** 181-187

Walther D, Brunnemann R, Selbig J (2007) The regulatory code for transcriptional response diversity and its relation to genome structural properties in A. thaliana. *PLoS Genetics* **3:** e11-e11

Wang J, Delabie J, Aasheim HC, Smeland E, Myklebost O (2004) Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Bioinformatics* **3:** 36-36

Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JDJ, Bertin N, Chung S, Vidal M, Gerstein M (2004) Annotation transfer between genomes: protein?protein interologs and protein?DNA regulogs. *Genome Research* **14:** 1107-1118

Zhang X, Shiu S, Cal A, Borevitz JO (2008) Global analysis of genetic, epigenetic and transcriptional polymorphisms in Arabidopsis thaliana using whole genome tiling arrays. *PLoS Genetics* **4:** e1000032-e1000032

Zhou X, Su Z (2007) EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics* **8:** 246-246

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiology* **136:** 2621-2632

3. Discovering regulatory programs underlying drought response in Arabidopsis

Arjun Krishnan, Amal Harb, Andy Pereira

3.1. Abstract

Drought is one of the most widespread of environmental stresses that has a large impact on crop growth and yield. The response of plants to drought stress is extremely complex and spans several orders of magnitude in time and space, causing system-wide changes that comprise of protective responses and adverse reactions. To understand this system-wide response, we performed gene expression profiling of Arabidopsis plants subjected to a controlled, sub-lethal, moderate drought (mDr) treatment similar to the stress encountered by crops in the field. In addition, expression profiles were also obtained from plants under a progressive drought (pDr) wilting treatment, along with publically available acute dehydration drought profiles to contrast with mDr. Computational analysis of differential expression was followed by detailed comparison across treatments at the levels of individual genes and regulated biological processes. *De novo* cis-regulatory element (CRE) analysis was then carried out to identify potential motifs that mediate the observed changes in gene expression. Transcriptional regulatory programs underlying drought-regulated gene expression under the various treatments were then mined based on significant associations between CREs and biological processes. Gene expression analysis led to the identification of a pDr induced MYB transcription factor gene, *GROWTH REGULATION UNDER DROUGHT 1 (GRD1)*, knockout of which confers drought sensitivity. Expression profile of drought stressed *grd1* mutant was obtained and compared systematically to that of WT to identify putative genes and processes that could mediate drought-regulated gene expression giving rise to sensitivity.

3.2. Introduction

Abiotic stresses are recognized as the primary causes of crop loss worldwide (Boyer, 1982). Drought is a major environmental stress problem and is expected to increase with climate change. Water is an increasingly scarce resource given current and future human population and

societal needs, putting an emphasis on sustainable water use. Plants have evolved specific acclimation and adaptation mechanisms to respond and survive short- and long-term environmental stresses. The ability to constantly sense and adapt to environmental changes to maintain cellular functions (homeostasis) is enabled by a complex network of genes (and their products) that needs to be unraveled (Shinozaki & Yamaguchi-Shinozaki, 2007).

Following the post-genomic wave, genome-wide expression profiling has been a valuable tool for determining an important facet of the complex response – transcriptional regulation of gene expression – during drought stress in plants, especially Arabidopsis (Kreps et al, 2002; Seki et al, 2002; Shinozaki et al, 2003). This is assuming that gene expression changes observed on a microarray are primarily due to transcriptional regulation (and not other modes of regulation including small RNA-mediated regulation). Extensive transcriptome data on response under dehydration/acute-drought (aDr) (Kilian et al, 2007) and progressive soil water deficit (pDr) (Huang et al, 2008) have provided insights into gene regulation under drought in terms of perturbed individual genes and associated biological processes. However, while aDr is an unnatural laboratory treatment, pDr pushes the plant towards extreme stress from where they are unable recover. Therefore, simulating field-like conditions of drought stress – a short period of drought from which tolerant plants can manage to survive and complete their growth cycle – will provide a better understanding of drought acclimation process and resistance mechanisms. To achieve this, Arabidopsis plants were subjected to a controlled, sub-lethal, moderate drought (mDr) treatment and gene expressions profiled 1 day and 10 days after treatment. In the work presented here, we analyze these expression profiles and compare them with gene expression under pDr and aDr to tease out the transcriptional regulatory changes that are common and specific to the different drought treatments.

Several large-scale integrative analyses have been performed on stress gene expression data to unveil sets of transcriptionally similarly regulated (co-responsive) genes and the biological processes they participate in. Hierarchical (Huang et al, 2008), probabilistic (Supper et al, 2007) and fuzzy k-means (Ma & Bohnert, 2007) clustering approaches have been used to group genes based on similar expression across different stresses, and then annotate the groups with the biological processes they participate in. Taking a simpler approach, starting with sets of

differentially expressed genes, biological processes perturbed in response to stress have been charted out for distinct stresses (Walther et al, 2007) or variants of broadly the same stress (Hannah et al, 2005). In the current work, we take the second approach for characterizing the processes and pathways regulated in mDr, pDr and aDr response. Once we understand which processes are transcriptionally regulated, the next question is to gather clues about how the regulation is brought about.

Regulation of gene expression is governed, in part, by the specific transcription factor binding motifs in the promoter regions of target genes. These cis-regulatory motifs in association with their transcription factors (TFs) are key in determining the cellular transcriptional response diversity – the spectrum of gene expression states elicited by different environmental stresses. Therefore, it would be of interest to identify the regulatory motifs that coordinate the observed gene expression changes under drought. In plants, a number of motifs related to stress response have been discovered by identifying distinct motifs associated with responses to specific treatments. Among these motifs, those mediating response to light, osmotic and cold stress treatments have been analyzed most intensely (Jiao et al, 2007; Yamaguchi-Shinozaki & Shinozaki, 2005). Databases dedicated to plant promoter motifs have also been established (Davuluri et al, 2003; Higo et al, 1999) based on motif identification in single or, at most, a few genes. All the gene expression studies mentioned previously have therefore used data from these databases to map 'known' motifs to upstream sequences of genes and find motifs 'enriched' among the sets of co-responsive genes of interest.

Here we present a comprehensive analysis of gene expression under moderate (mDr), progressive (pDr) and acute dehydration drought (aDr) by improving upon previous methods. First, we compare the response of the plant to these three stress regimens at the level of individual genes. Second, we identify and compare biological processes involved in the different drought stress responses. Third, we perform *de novo* cis-regulatory motif discovery on the sets of co-responsive genes. This analysis has the potential to find short degenerate DNA sequences in the upstream regions of genes from scratch that are informative about the genes' expression patterns, which therefore could lead to discovery of correct/general forms of known motifs or completely novel motifs. Finally, we associate perturbed biological processes to specific

regulatory motifs using enrichment analysis within the context of the perturbation, thus revealing the transcriptional regulatory programs underlying drought response in Arabidopsis.

Furthermore, based on the above analysis we identified a MYB transcription factor gene *Grd1* responsive to pDr, which when knocked out gives drought sensitivity. This mutant (*grd1*) was also profiled for gene expression changes under drought stress for further study. Therefore, to conclude, we describe a similar analysis based on regulatory programs that helps elucidate the responses of the mutant in relation to the drought response in the wild-type (WT) plant.

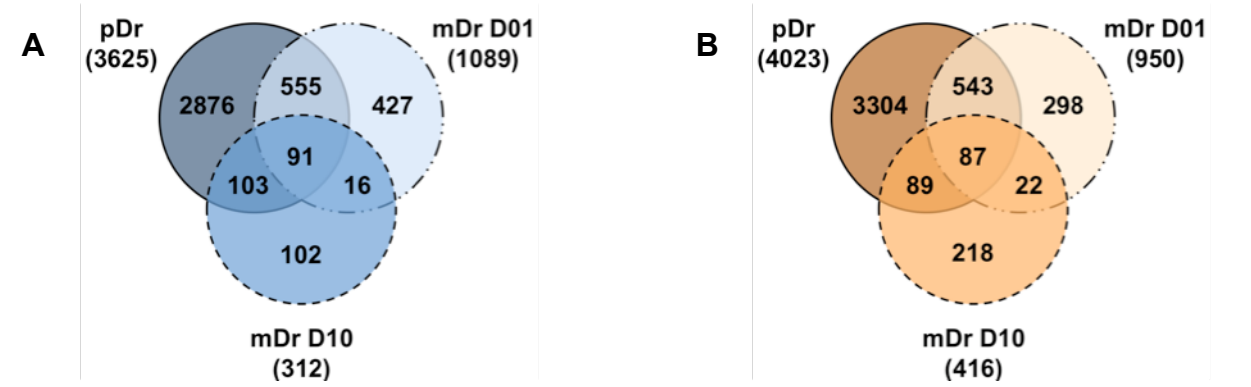
3.3. Results

3.3.1. Gene expression profiles of moderate, progressive and acute drought in Arabidopsis

In order to understand the global effects of drought stress on gene expression, microarrays were used to profile genome-scale gene expression levels under moderate drought (mDr; Day1 and Day10) and progressive drought (pDr) conditions and their corresponding controls in samples from young leaves. Analysis of differential expression based a custom gene-centric probeset annotation showed that a large number of genes (2039) were significantly perturbed very early (Day1) in response to mDr. In contrast, after a prolonged moderate drought treatment (Day10), a far less number of genes (728) were differentially expressed. Compared to the two mDr treatments, severe effects of drought on gene expression were revealed by the response to pDr (wilting): 7648 differentially expressed (DE) genes, about 30% of the genes in the genome.

Comparison of the three drought treatments – mDr-Day1, mDr-Day10 and pDr – was carried out first at the gene-level (Fig. 3.1A and B). A common set of 178 genes responded to mDr and pDr treatments (91 up- and 87 down-regulated), while 1083 (545 up- and 538-down regulated) genes were specific to mDr. Another observation is that the Jaccard's coefficient ([No. genes in the intersection]/[No. genes in the union]) between mDr Day1 and pDr is ~15.8%, that between mDr Day10 and pDr is ~5% and that between mDr Day01 and mDr Day10 is ~8%, indicating that mDr Day 01 is more similar to pDr than mDr Day01. Most of our knowledge of cellular drought response is based on dehydration (acute drought; aDr) and/or osmotic treatments of plants under laboratory conditions (Kilian et al, 2007). Hence, we compared our mDr and pDr response genes

to those responding to aDr at various time points ranging from 15min to 24h (Fig. 3.1C). Apparently, there is very little overlap between mDr and aDr time course response. In fact, while there is some similarity between mDr response in Day1 and later time points of aDr, mDr Day10 seems to be showing the inverse of aDr response with the genes down-regulated in mDr-Day10 are up-regulated in earlier time points (30min to 3h) of aDr. This comparison lends support to the hypothesis that the moderate drought treatment elicits a very different response in the plant as against a drastic treatment like aDr. On the other hand, there is a reasonably high amount of overlap between pDr and aDr with maximum agreement occurring at 6h and 12h of aDr for up- and down-regulated genes. This similarity between the aDr and pDr (where the plant is just about wilting) again is in agreement with the above hypothesis. An aspect of aDr response that stands out is the gene response at the earliest time point (15min): genes induced early by aDr are among those that are repressed by mDr (Day1) and pDr, and genes repressed early by aDr are induced by pDr, showing that there is a possibility for activation of a protective immediately after the sudden dehydration treatment.



C

			aDr	Up							Down						
				0.25h	0.5h	1h	3h	6h	12h	24h	0.25h	0.5h	1h	3h	6h	12h	24h
				(1871)	(1819)	(1344)	(1201)	(1431)	(97)	(557)	(923)	(1739)	(1865)	(1202)	(1443)	(61)	(1567)
mDr Day01	Up	(1089)		0.051	0.198	0.172	0.163	0.211	0.320	0.156	0.139	0.108	0.053	0.034	0.028	0.016	0.041
	Down	(950)		0.273	0.107	0.075	0.033	0.028	0.021	0.045	0.061	0.199	0.245	0.205	0.266	0.213	0.239
mDr Day10	Up	(312)		0.061	0.186	0.170	0.109	0.186	0.072	0.096	0.170	0.179	0.109	0.093	0.019	0.016	0.016
	Down	(416)		0.077	0.267	0.305	0.204	0.147	0.082	0.041	0.154	0.200	0.113	0.067	0.077	0.049	0.055
pDr	Up	(3625)		0.119	0.304	0.371	0.380	0.516	0.433	0.372	0.380	0.212	0.135	0.090	0.033	0.148	0.119
	Down	(4023)		0.468	0.204	0.159	0.190	0.089	0.093	0.171	0.198	0.364	0.429	0.556	0.639	0.623	0.489

Overlap = $|A \cap B| / min(|A|, |B|)$ >0.2 >0.3 >0.4 >0.5 >0.6

Figure 3.1: Gene expression profiles under moderate (mDr) and progressive (pDr) drought.

Venn diagrams comparing up- (**A**) and down-regulated (**B**) genes of moderate drought 1 day (mDr D01), 10 day (mDr Day10) and pDr wilting treatments. Total numbers of genes for all genesets are indicated in brackets. (**C**) Table of overlap in gene-regulation between the drought treatments presented in this study – mDr and pDr – and published dehydration acute drought (aDr) (Kilian et al, 2007). Overlap between a pair of drought-regulated gene sets (each cell in the table) is measured as the number of common genes between the sets divided by the number of genes in the smaller of the two sets. The color-scale from light to dark green is used as an indicator of the degree of overlap from small to large. As used here and in the rest of the figures, shades of blue are used to indicate up-regulation and shades of yellow are used to indicate down-regulation.

Overall, mDr, pDr and aDr thus offer a very diverse set of plant responses to different drought treatments and are hence useful in gaining a broad picture of drought-regulated gene expression.

3.3.2. Biological processes up- and down-regulated across drought treatments

To unravel the biological processes and pathways that are perturbed by the various drought treatments, drought response gene sets from each of mDr, pDr and aDr were functionally characterized using enrichment analysis of gene sets described based on Gene Ontology (GO) (Ashburner et al, 2000) biological process annotation terms. As expected, at the core of the induced gene expression across drought treatments are the stress response genes: genes responding to water deprivation, abscisic acid (ABA) stimulus, and salt, osmotic and cold stresses (Fig. 3.2A). Expression dynamics of several of these genes has been verified using qRT-PCR (data not shown). Prominent fundamental processes repressed by the different drought treatments include DNA packaging, nucleobase biosynthesis, several related to protein synthesis (rRNA processing, ribosome biogenesis, tRNA modification, amino acid biosynthesis, and translational initiation and elongation), protein folding and energy production.

It was observed that there exists a strong induction of jasmonic acid (JA) related processes like ‘response to JA stimulus’, ‘response to wounding’, and JA biosynthesis and signaling from 30min to 6h of aDr. In contrast, the same processes were repressed under mDr-Day10. To verify the biological relevance of this observation, JA signaling mutants *coi1* and *jin1* were tested under mDr and were shown to display significant drought resistance at the end of mDr treatment (Day10) compared to control plants (data not shown).

Among the time profiles in aDr, the response mounting immediately (15min) after dehydration offers a very interesting perspective on the early response to severe stress that is consistently inverse of the response in any other time point or drought treatment. The earlier gene-level

comparison (Fig. 3.1C) had given clues in this direction. At the process-level, the first observation is that several of the previously known stress response genes induced by the various drought treatments are not differentially expressed at this early time point. Second, many drought-repressed processes, especially those related to nucleotide biosynthesis, protein synthesis and protein folding, are strongly induced at 15min. Interestingly, transcription in general is also strongly up-regulated (only) at this time point supporting the emerging hypothesis that there might be a lot of constructive cellular response at this early stage.

Processes unique to mDr and pDr treatments (not observed under severe dehydration in aDr) are presented in Fig. 3.2B. Of particular interest here are photosynthesis (and related processes) and RNA splicing that are down- and up-regulated only under pDr. The distinctive reaction of the plant to mDr-Day1, early response to a moderate stress, is the activation of plant cell wall modification genes that underlie cell growth, which are in-fact down-regulated by pDr. This aspect of mDr response was pursued experimentally and confirmed (data not shown).

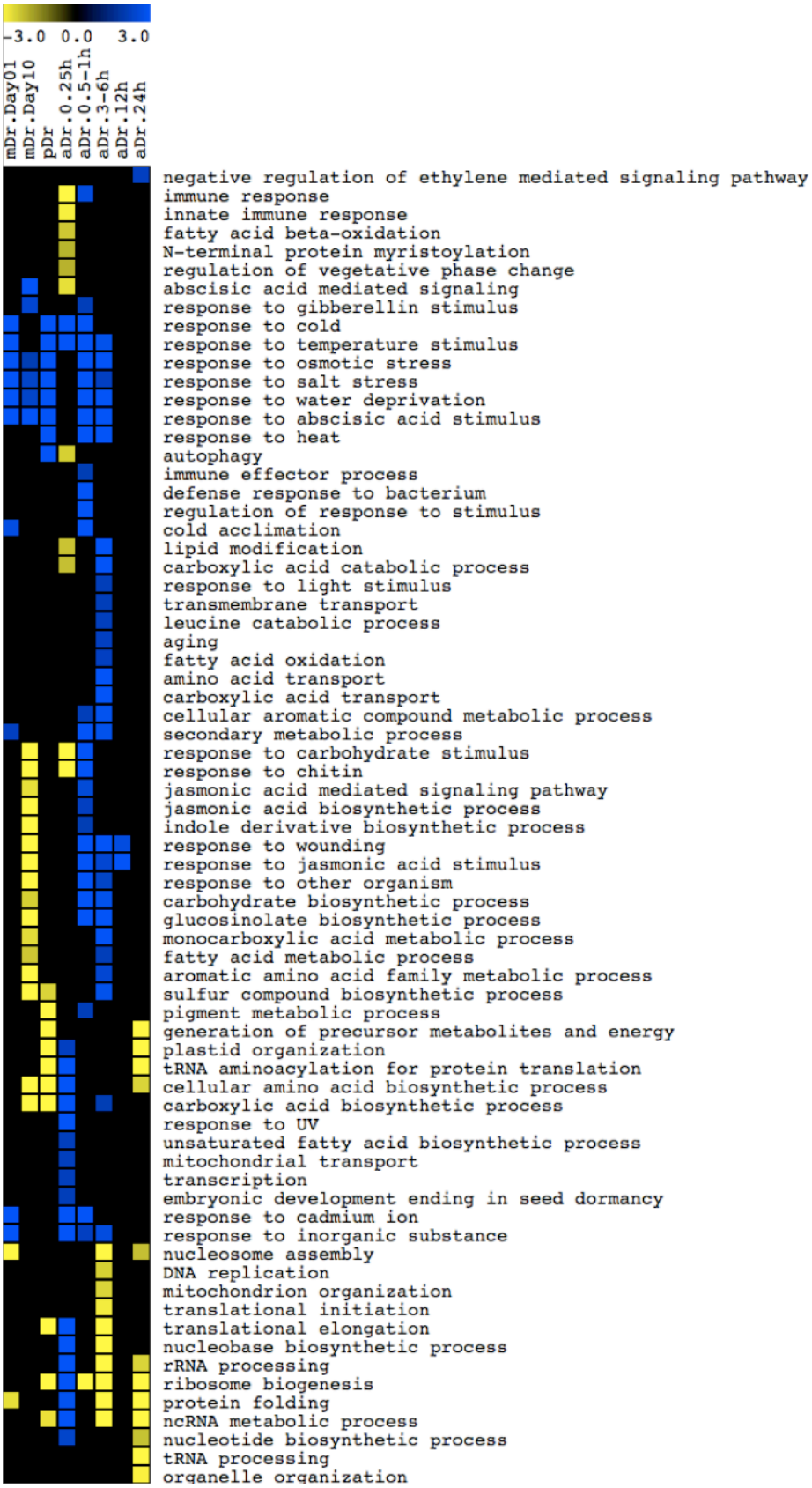
Together, these perturbed biological processes present a spectrum of drought response in plants and raises even more critical questions: what transcriptional regulation underlies the differential expression of these genes and processes under the different drought treatments?

3.3.3. Cis-regulatory elements coordinating gene expression changes under drought

In order to glean some clues about the components of the regulatory networks that mediate the transcriptional regulation of the drought response genes, we turned to *de novo* identification of cis-regulatory elements (CREs). This was a feasible strategy relying on two well-known sources of information, namely gene-regulation under a condition (any drought treatment) and upstream sequences of all the genes in the genome. In addition, since known regulatory motifs are based on knowledge from one or a few genes, there is a large scope for revising these motifs to make them more complete and accommodate degeneracy, especially in the context of a conditional response (drought stress). And, finally, there are certainly several novel motifs that need to be discovered that mediate transcriptional regulations inexplicable based on current knowledge of regulatory motifs.

We devised a CRE-discovery pipeline using the de-novo motif discovery tool, FIRE (Elemento et al, 2007). We subsequently compared the newly identified motifs to known cis-elements in the PLACE database of known cis-regulatory elements (Higo et al, 1999) using STAMP (Mahony & Benos, 2007). Since there are several different treatments involved, in order to systematically clarify the motifs identified for each of the response gene sets, we compared all the discovered motifs to each other, again using STAMP and built a similarity tree. Sequence logos of CREs of interest were then produced using WebLogo (Crooks et al, 2004). Applying this pipeline to mDr, pDr and aDr (further separated into up- and down-regulated) gene sets, led to the identification of several known and novel CREs. The discovered motifs, the drought response gene sets they were discovered in, their comparison to each other and comparison to known motifs are presented in Figure 3. Motifs discussed below are referred to by the name of the most similar 'known' CRE (see the 'PLACE motifs' table in Figure 3 for the key).

A


-3.0 0.0 3.0
mDr.Day01
mDr.Day10
pDr
aDr.0.25h
aDr.0.5-1h
aDr.3-6h
aDr.12h
aDr.24h
negative regulation of ethylene mediated signaling pathway
immune response
innate immune response
fatty acid beta-oxidation
N-terminal protein myristoylation
regulation of vegetative phase change
abscisic acid mediated signaling
response to gibberellin stimulus
response to cold
response to temperature stimulus
response to osmotic stress
response to salt stress
response to water deprivation
response to abscisic acid stimulus
response to heat
autophagy
immune effector process
defense response to bacterium
regulation of response to stimulus
cold acclimation
lipid modification
carboxylic acid catabolic process
response to light stimulus
transmembrane transport
leucine catabolic process
aging
fatty acid oxidation
amino acid transport
carboxylic acid transport
cellular aromatic compound metabolic process
secondary metabolic process
response to carbohydrate stimulus
response to chitin
jasmonic acid mediated signaling pathway
jasmonic acid biosynthetic process
indole derivative biosynthetic process
response to wounding
response to jasmonic acid stimulus
response to other organism
carbohydrate biosynthetic process
glucosinolate biosynthetic process
monocarboxylic acid metabolic process
fatty acid metabolic process
aromatic amino acid family metabolic process
sulfur compound biosynthetic process
pigment metabolic process
generation of precursor metabolites and energy
plastid organization
tRNA aminoacylation for protein translation
cellular amino acid biosynthetic process
carboxylic acid biosynthetic process
response to UV
unsaturated fatty acid biosynthetic process
mitochondrial transport
transcription
embryonic development ending in seed dormancy
response to cadmium ion
response to inorganic substance
nucleosome assembly
DNA replication
mitochondrion organization
translational initiation
translational elongation
nucleobase biosynthetic process
rRNA processing
ribosome biogenesis
protein folding
ncRNA metabolic process
nucleotide biosynthetic process
tRNA processing
organelle organization

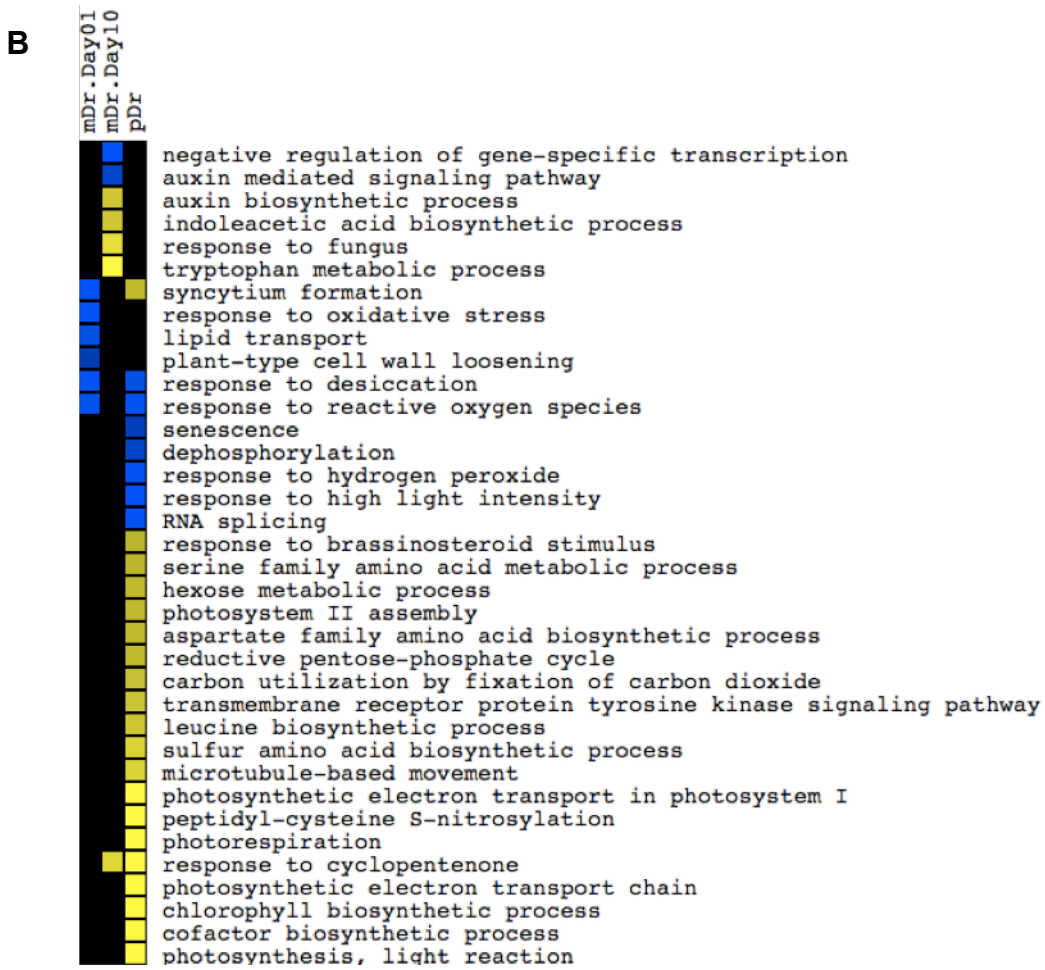


Figure 3.2: Functions and processes common and specific to various drought stress treatments and time-points. These are defined broadly based on Gene Ontology (GO) biological process (BP) annotations of Arabidopsis genes defined by TAIR. GO BP terms of interest (rows) were identified by analysis of enrichment of the set of genes annotated with a given GO BP term in each of the (up and down) drought-regulated gene sets (columns). Statistical significance of enrichment was calculated using the hypergeometric test and terms with *q-value* <0.01 in at least one of the treatments were retained. The enrichment of each term is indicated using a yellow-blue color scheme with the intensity of the color representing the degree of enrichment (measured as the $-log_{10}$*[q-value]*) and the color indicating enrichment among the down- or up-regulated genes in each treatment. Black indicates no enrichment. (**A**) GO BP terms enriched among aDr-regulated genes and their enrichment among the mDr and pDr genes. (**B**) GO BP terms specific to moderate and progressive drought treatments.

The CRE that was most significant among the genes up-regulated in mDr, pDr and several aDr time points was one highly similar to the experimentally identified ACGT-containing ABRE-motif ACGTG(G/T)C (Fig. 3.3 and 3.4) (Hattori et al, 2002). Indeed, in the pDr-ABRE, the (G/T) degeneracy is exactly preserved at position 6. The strict T at this position in the mDr-Day1-ABRE suggests that the two variants of the ABRE motif, one with G and the other with T at position 6, might have slightly different roles or be bound by different bZIP TFs belonging to the same family with slightly different specificities. Interestingly, two element very similar to the

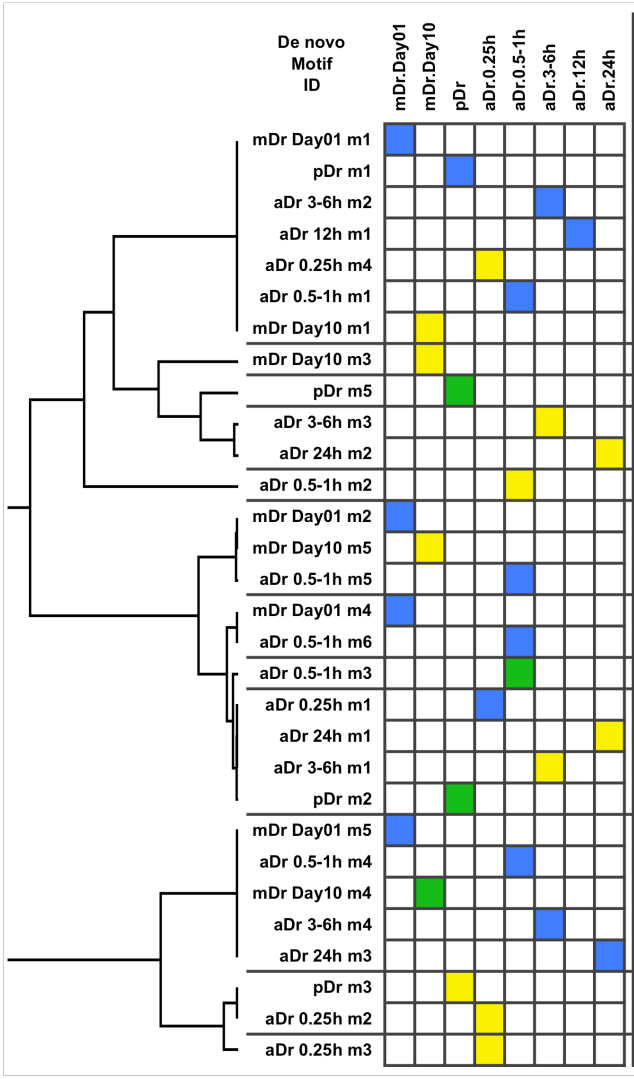
ABRE – A(A/C)(A/C)(A/G)CGTG and (A/C/T)(A/C)ACGCGT(A/C/T) – were found among genes down-regulated in mDr-Day10 and aDr-15min. Both these subtly different elements are indeed more similar to coupling element 3 (CE3) that is known to be part of the necessary and sufficient ABA-responsive cis-regulatory module (Shen et al, 2004). The role of CE3 in contrast to ABRE is to be explored further.

Another motif associated with up-regulated gene expression in mDr-Day10 and aDr-30min-1h, and down-regulation in mDr-Day10 was one highly similar to the DRE/CRT-motif (A/G)CCGAC. In support of the co-occurrence of DRE/CRT and ABRE in the same set of genes, the two elements have been found to be interdependent in the dehydration-responsive expression of the *rd29A* gene in Arabidopsis (Narusaka et al, 2003).

Among the down-regulated genes, the Ibox motif (Giuliano et al, 1988) was discovered in the pDr-regulated genes where we observe strong repression of the photosynthetic machinery, thus suggesting a cross-talk between the light-signaling and stress-response pathways. However, a motif was also discovered among genes down-regulated in aDr-15min where its role is unclear since photosynthetic genes are not regulated.

Motifs similar to the site II motif, known to be coordinating transcriptional regulation of ribosomal protein genes (Tremousaygue et al, 2003), were also found among the genes down-regulated in aDr-3-6h and aDr-24h, and up-regulated in aDr-15min, precisely consistent with the pattern of regulation of genes related to ribosome biogenesis in these time points of aDr.

Yet another classical stress-response was found in early mDr, comprising slowing down of protein folding in the ER that triggers the unfolded protein response (Martinez & Chrispeels, 2003), suggested by the identification of the UPRE-like element among genes up-regulated in mDr-Day1. In agreement, the category ‘protein folding’ was enriched among the down-regulated genes in mDr-Day1. Moreover, several genes involved in cell elongation and division are down-regulated by UPR (Martinez & Chrispeels, 2003), genes that are up-regulated in mDr-Day1.



De novo Motif ID	De novo Motif Sequence	Z-score	PLACE motifs					
			ID	Sequence	E-value	ID	Sequence	E-value
mDr Day01 m1	ACGTGTM	63.807	GADOWNAT	ACGTGTC	9.88E-11	ABREMOTIFAOSOSEM	TACGTGTC	8.77E-10
pDr m1	ACGTGKC	172.805	ACGTABREMOTIFA2OSEM	ACGTGKC	1.05E-12	BOXIIPCCHS	ACGTGGC	9.88E-11
aDr 3-6h m2	CACGTKY	64.673	CACGTGMOTIF	CACGTG	4.99E-08	ABRE2HVA22	GCCACGTACA	3.61E-07
aDr 12h m1	WDHACGTGD	19.809	ABRE3HVA22	GCCACGTACA	1.94E-06			
aDr 0.25h m4	HMACGCGTH	13.328	CE3OSOSEM	AACGCGTGTC	1.52E-08	ABADESI2	GGACGCGTGGC	5.31E-08
aDr 0.5-1h m1	DMCRCGTDK	114.108	GBOXLERBCS	MCACGTGGC	8.83E-08	ABRERATCAL	MACGYGB	1.54E-07
mDr Day10 m1	AMMRCGTG	50.285	CE3OSOSEM	AACGCGTGTC	6.79E-08	GBOXPC	ACCACGTGGC	6.79E-08
mDr Day10 m3	KMCAGCTVW	31.884						
pDr m5	HATCCGDAD	21.235	REBETALGLHCB21	CGGATA	3.47E-06			
aDr 3-6h m3	AWVCCGGB	20.85						
aDr 24h m2	VNAACCKGW	25.062						
aDr 0.5-1h m2	RMRCCCTR	29.999	UP2ATMSD	AAACCCTA	1.08E-07	TELOBOXATEEF1AA1	AAACCCTAA	5.45E-07
mDr Day01 m2	DGCCGACH	24.051	DRECRTCOREAT	RCCGAC	1.81E-07			
mDr Day10 m5	VCCGACMWH	22.959	DRECRTCOREAT	RCCGAC	4.95E-07			
aDr 0.5-1h m5	YGACCGAY	17.095	DRE2COREZMRAB17	ACCGAC	1.81E-07			
mDr Day01 m4	DTRGTCCAM	17.149	UPRE1AT	ATTGGTCCACG	8.06E-09			
aDr 0.5-1h m6	RGTCAAC	14.436	WBBOXPCWRKY1	TTTGACY	2.44E-08			
aDr 0.5-1h m3	BHBGGTCCH	25.777						
aDr 0.25h m1	MRGCCCAD	87.022	SITEIIATCYTC	TGGGCY	3.65E-09			
aDr 24h m1	HMRGCCCAH	79.632	SITEIIATCYTC	TGGGCY	1.18E-08			
aDr 3-6h m1	HARGCCCH	83.517	SITEIIATCYTC	TGGGCY	1.39E-06			
pDr m2	RTGGVCC	80.676	GCBP2ZMGAPC4	GTGGGCCCG	1.03E-06	UPRE1AT	ATTGGTCCACG	4.33E-06
mDr Day01 m5	AAAAAATD	15.991	PYRIMIDINEBOXHVEPB1	TTTTTTCC	1.16E-06			
aDr 0.5-1h m4	AAAAAAG	24.746	-314MOTIFZMSBE1	ACATA{4}TA{7}GGCA	2.06E-08			
mDr Day10 m4	AAAATATH	29.35	EVENINGAT	AAAATATCT	9.68E-11	LECPLEACS2	TAAAATAT	2.57E-10
aDr 3-6h m4	AAAATAAW	14.327	TATABOX5	TTATTT	3.65E-09	-314MOTIFZMSBE1	ACATA{4}TA{7}GGCA	8.47E-09
aDr 24h m3	HTAAAATAH	22.938	LECPLEACS2	TAAAATAT	9.68E-11	-314MOTIFZMSBE1	ACATA{4}TA{7}GGCA	1.78E-07
pDr m3	VBCTTATCY	28.036	IBOXCORENT	GATAAGR	4.19E-09			
aDr 0.25h m2	CTTATCCH	31.054	IBOX	GATAAG	3.65E-09			
aDr 0.25h m3	YATAATTA	15.357	HDZIP2ATATHB2	TAATMATTA	7.01E-08			

Figure 3.3: Cis-regulatory elements identified in the upstream regions of drought-regulated genes from the various mDr, pDr and aDr treatments.
Each element identified along the rows was identified using *de novo* motif discovery to identify short degenerate DNA sequences whose presence or absence in the 1Kb upstream regions of genes is highly informative about the expression of the given gene set (e.g. up-regulated genes in mDr Day 01) given the background distribution of the sequence in the upstream sequences of all the genes in the genome. The 30 motifs thus identified were compared to each other to reveal that groups of very similar underlying motifs were discovered independently based on the gene expression under different treatments, represented using a dendogram on the left extreme. The colored matrix indicates which motifs were identified using genes regulated in which drought treatment, again with yellow indicating down-regulation and blue up-regulation. Motifs informative about up- and down-regulation together are indicated by green. In the adjoining table, the sequence of the *de novo* motifs are given in the nucleotide IUPAC nomenclature along with the Z-score of the information value of the motif reflecting how far the observed value is, in number of standard deviations, from the average random information (see *Methods*). These motifs were then compared to known cis-elements in the PLACE database using the STAMP web server. Known elements with significant match to each *de novo* motif are presented in the 'PLACE motifs' table in the form of the database ID, DNA sequence and E-value of sequence match with the *de novo* motif. Motifs with no match to any known element are novel putative regulatory elements.

And, finally, three distinct novel motifs have been discovered among genes down-regulated in mDr-Day10, down-regulated in aDr3-6h and aDr-24h, and up- and down-regulated in aDr-1-3h. It is clear that there is a diverse set of motifs, both known and novel, that might coordinate gene regulation in response to various drought treatments. It is therefore important to scope-down on the exact processes that each of the above mentioned cis-elements could potentially regulate under drought, to the best resolution possible.

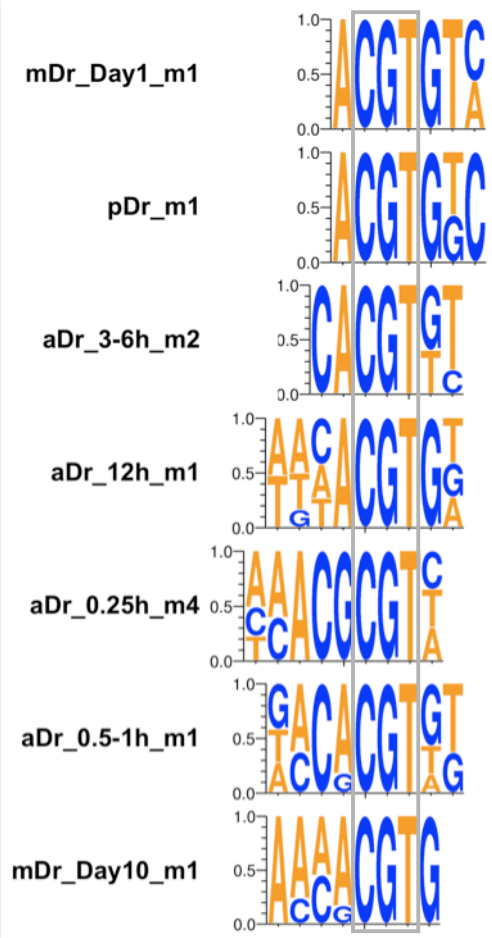


Figure 3.4: Sequence logos of ABRE-like cis-regulatory motifs discovered in the different drought-regulated gene sets corresponding to the first group of motifs presented at the top of Figure 3.
The height of a nucleotide in a particular position in each logo corresponds to the fraction of genes with the corresponding expression pattern (e.g. pDr_Up) that contain that nucleotide in that position. The logos are arranged so as to align the central 'CGT' core, marked with the grey box. Motifs mDr_Day1_m1, pDr_m1, aDr_3-6h_m2, aDr_m1 and aDr_0.5-1h_m1 are identified among the up-regulated genes in the corresponding treatments, while motifs aDr_0.25h_m4 and mDr_Day10_m1 are identified among the down-regulated genes. Although it appears that all the motifs are similar to the ABRE motif ACGTG[GT]C, and that it is involved in both up- and down-regulation, careful inspection of the table mapping to known motifs in Figure 3 and the sequence logos here shows that the motifs present among the down-regulated genes (aDr_0.25h_m4 and mDr_Day10_m1) are actually closer to the coupling element AACGCGTGTC than to ABRE, thus clearly differentiating them from the other motifs more closely similar to the ABRE motif.

3.3.4. Transcriptional regulatory programs associating regulatory elements to perturbed biological processes

Once the perturbed processes/pathways and putative CREs have been identified, rarely have studies tied these two pieces together to reveal putative transcriptional regulatory programs: CREs (individually or in combinations) that could account for cis-regulatory modules mediating

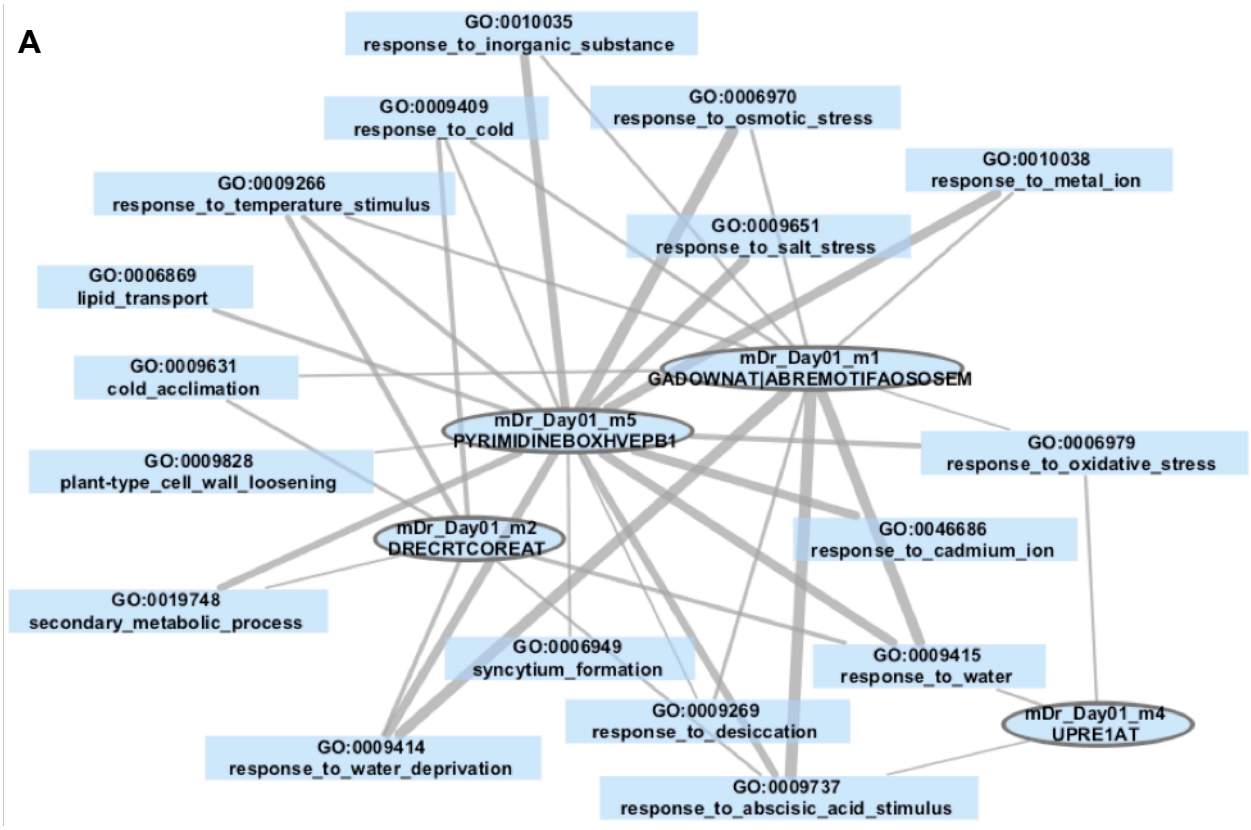
the regulation of a specific or groups of related processes. Here, we make an effort to fill this gap and thereby describe CRE-process associations that are relevant to drought response.

Broadly, within the context of each drought-regulated gene set, when genes containing a CRE relevant to that gene set significantly overlapped with genes annotated to a relevant GO BP term, then the CRE-GO_BP pair was considered to be relevant and biologically associated. Several such pairs were identified for the various drought treatments, time points and patterns of regulation and are presented in Figure 5 and 6 as in the form of a graph connecting CRE nodes with GO BP term nodes whenever the pair is significant.

The first observation is that CREs work in a combinatorial manner as expected and coordinate the regulation of groups of processes. For example, in mDr-Day1, the ABRE- and DRE-like motifs are associated with and could cooperatively regulate the core set of stress response genes (Fig. 3.5A) as observed previously (Narusaka et al, 2003). Among the other CREs in mDr-Day1, we observed that the pyrimidine-box-like motif is also connected to the stress-response processes. However, these motifs are typically found among gibberellic acid (GA) regulated gene induction (Mena et al, 2002), and GA is known to act antagonistically to ABA (Piskurewicz et al, 2009) (as also evident from the ABRE-like motif also being similar to GADOWNAT motif), making the association of this motif with the up-regulated genes unclear. Nevertheless, this motif is seems to underlie the up-regulation of cell wall-related genes in mDr-Day1.

In the case of mDr-Day10, we found that the CE3-like motif that present among the down-regulated genes is connected with the down-regulation JA-response genes. Cyclopentenones are related to JA biosynthesis and play a role in defense response in the absence of JA (Stintzi et al, 2001), suggesting that the enrichment of this process along with JA-response is plausible. It is of interest here because the novel motif (G/T)(A/C)CAGCT(A/C/G)(A/T) (mDr_Day10_m3) is associated with this process.

The possible roles of ABRE-like motif and Ibox-motif are clear with respect to pDr response and the expected connections between these CREs with the up- and down-regulated processes have been recovered (Fig. 3.5C and D).


A
GO:0010035
response_to_inorganic_substance
GO:0009409
response_to_cold
GO:0006970
response_to_osmotic_stress
GO:0010038
response_to_metal_ion
GO:0009266
response_to_temperature_stimulus
GO:0009651
response_to_salt_stress
GO:0006869
lipid_transport
GO:0009631
cold_acclimation
mDr_Day01_m1
GADOWNAT|ABREMOTIFAOSOSEM
mDr_Day01_m5
PYRIMIDINEBOXHVEPB1
GO:0006979
response_to_oxidative_stress
GO:0009828
plant-type_cell_wall_loosening
mDr_Day01_m2
DRECRTCOREAT
GO:0046686
response_to_cadmium_ion
GO:0019748
secondary_metabolic_process
GO:0006949
syncytium_formation
GO:0009415
response_to_water
GO:0009269
response_to_desiccation
mDr_Day01_m4
UPRE1AT
GO:0009414
response_to_water_deprivation
GO:0009737
response_to_abscisic_acid_stimulus
B
GO:0032787
monocarboxylic_acid_metabolic_process
GO:0009611
response_to_wounding
mDr_Day10_m3
mDr_Day10_m1
CE3OSOSEM|GBOXPC
GO:0010583
response_to_cyclopentenone
GO:0009753
response_to_jasmonic_acid_stimulus
GO:0009605
response_to_external_stimulus

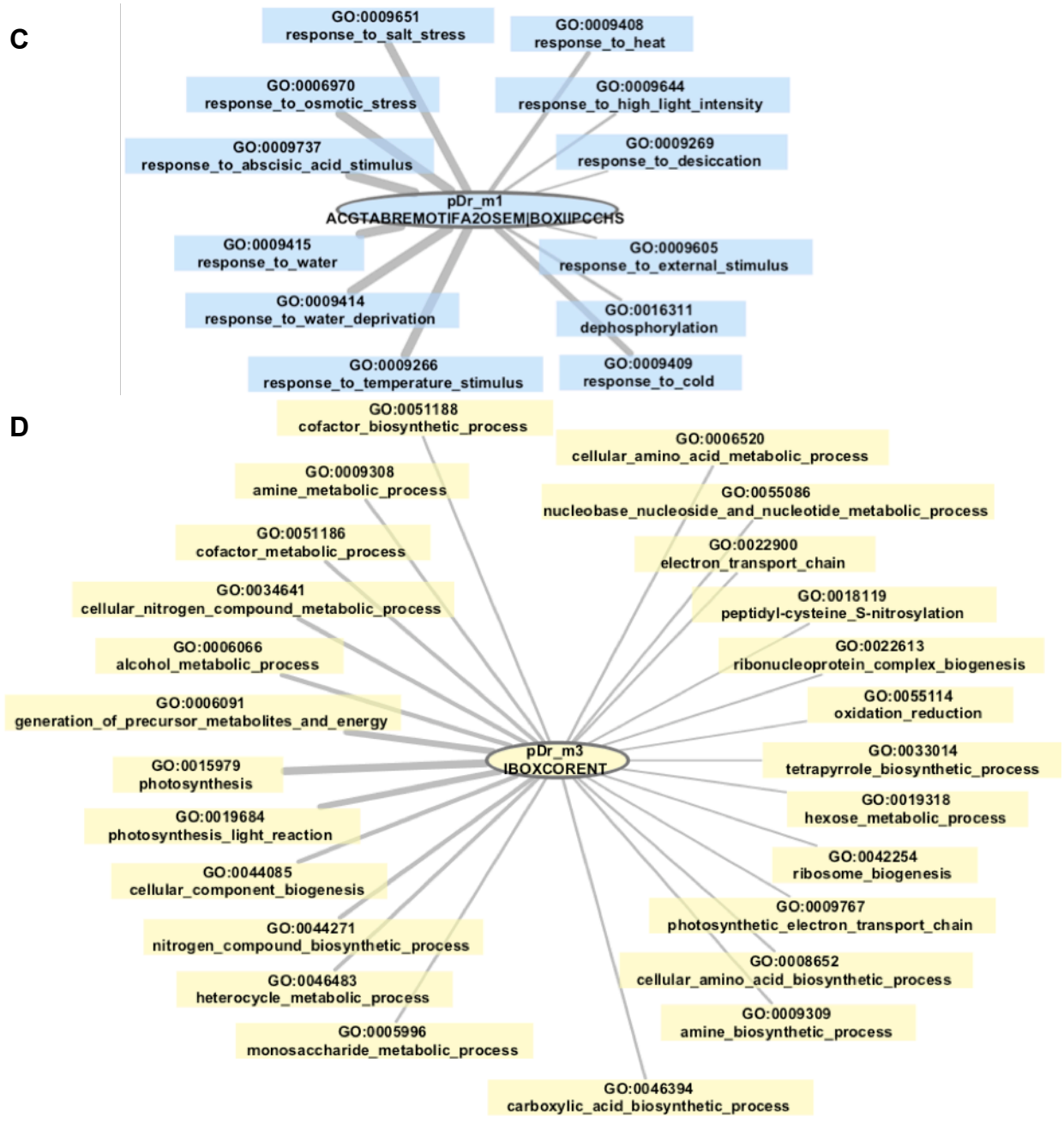


Figure 3.5: Transcriptional regulatory programs underlying mDr and pDr response that reveal connections between putative cis-regulatory elements and the functions of their target genes.
For each drought-responsive gene expression pattern (e.g. mDr Day1 up-regulation or pDr down-regulation), genes containing concerned putative cis-regulatory motifs were significantly enriched among genes annotated to different sets of concerned GO BP terms. Such associations are represented by a graph with edges connecting motifs (ovals) to GO BP terms (rectangles). In each case, the labels of the motifs represent the drought treatment under consideration and the color of the node represents the expression pattern, with yellow indicating down-regulation and blue up-regulation: (**A**) mDr Day 1 Up; (**B**) mDr Day10 Down; (**C**) pDr Up; and (**D**) pDr Down. Thickness of the edges corresponds to the level of significance of enrichment, measured as the $-log_{10}[q\text{-}value]$ (with the *q-value* of the hypergeometric test). GO BP terms presented here are among those in Figure 2. Motif labels correspond to their identifiers and annotations presented in Figure 3 that serves as the key.

Figure 3.6: Transcriptional regulatory programs (TRPs) underlying induction (**A**) and repression (**B**) 30min to1h post acute dehydration drought (aDr) reveal distinct regulatory modules each being a combination of cis-regulatory elements associated with groups of GO BPs.
These TRP and others identified for the rest of the phases of aDr-response are provided as Figures 3.S1-S4 here: http://treebeard.vbi.vt.edu/AK_Thesis/.

As an example of programs concerning aDr-response, the CRE-GO_BP associations for aDr-30min-1h have been presented here (Fig. 3.6). At least four distinct cis-regulatory modules

coordinate the up-regulation of different groups of processes: ABRE, DRE and minus314 motifs (Kim & Guiltinan, 1999) (similar to the pyrimidine box motif identified among the mDr-Day1 genes and working with ABRE and DRE motifs) with the stress response processes, minus314 motif with carbohydrate metabolism, ABRE and minus314 motifs with JA-response and signaling, 314minus and WBOX motifs with immune-response-related processes.

3.3.5. *GRD1* gene induced under progressive drought and required for growth under drought and salt stress

Based on the analysis of pDr-responsive gene-regulation, we identified a MYB TF gene GRD1 that was induced by pDr. Based on GO annotations, this gene also responds to salt, oxidative and ABA stimuli. Therefore, we hypothesized that this gene might play a role in drought response of the plant. To test this hypothesis, *grd1* knockout lines were subjected to progressive drought. Based on the measurement of a number of physiological parameters, we found that the *grd1* lines are sensitive to drought (data not shown). Hence, to gain some idea about the possible cellular changes in the *grd1* lines that contribute to its sensitivity, gene expression analysis of drought treated *grd1* plants was carried out. Below, we present the results from the large-scale analysis of the gene expression profile of *grd1* in comparison to that of WT.

Following differential expression analysis to identify drought-responsive genes in *grd1*, we found that 5748 genes were significantly regulated. Comparing these genes to those regulated in the WT under drought (Fig. 3.7A), we found discerned that although there is a large overlap between the drought-responsive genes in the WT and *grd1*, there we several genes belonging to sets of specific regulation (regions numerically labeled in Fig. 3.7A). To perform a systematic analysis, we first generated a hypothetical minimal model to explain the observed patterns of gene expression in the drought treated *grd1* lines compared to WT (Fig. 3.7B). Based on the facts that drought induces the expression of the GRD1 gene and the gene codes for a TF, the proposed model assumes that this gene (TF) mediates a part of the gene regulation under drought, and then accounts for the six major patterns of gene regulation observed in the *grd1*-WT comparison (Fig. 3.7A).

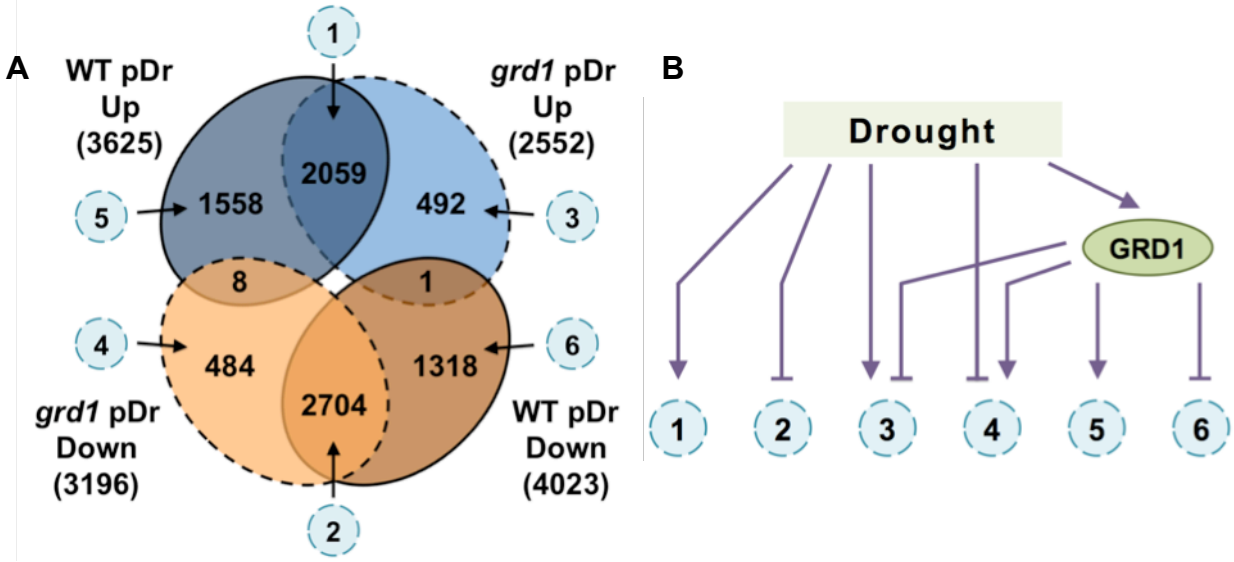


Figure 3.7: Comparison of gene expression patterns in WT and *grd1* mutant under progressive drought.
The six regions in the Venn diagram that contain considerable number of genes have been indicated using numerical labels from 1 through 6. (**B**) Hypothetical model for the role of *GRD1* in mediating drought-regulated gene expression. GRD1 is induced under drought and encodes a transcription factor (TF). Therefore, GRD1 could potentially mediate the regulation of some of the genes differentially expressed under drought. The proposed model accounts for the gene expression patterns observed in (**A**) by incorporating GRD1 into the drought-regulated gene expression. The sets of genes used here correspond to the six regions in the Venn diagram in (**A**). Sets 1 and 2 are sets of genes that are up- and down-regulated, respectively, under drought irrespective of the genotype (WT or *grd1*). Sets 5 and 6, on the other hand, are up- and down-regulated only in the WT but not in the *grd1* mutant, indicating that GRD1 potentially plays a role in drought-regulation of these sets of genes. Moreover, sets 3 and 4 are up- and down-regulated only in the *grd1* mutant suggesting that the absence of the gene causes drought induction and repression, which leads to the notion that the intact gene works antagonistically with drought in regulating the expression of these sets of genes. Prominent processes in these sets of *grd1*/pDr genes discussed in the text are related to jasmonic acid stimulus and signaling in set 3, and mRNA splicing in set 5.

We then separated all the drought-regulated genes in the WT and *grd1* into the six sets and performed CRE discovery and GO BP enrichment analyses followed by description of transcriptional regulatory programs. The end results of these analyses are presented in Figure 3.8 and follow the same format as Figures 5 and 6. The CREs identified here are detailed in Figure 3.S5. The common response, both up- and down-regulation, are as expected, and we therefore turn to the *differently* regulated gene sets, specifically sets 3 and 5. Genes in set 3 are those up-regulated only in *grd1* and not in the WT, suggesting that the absence of GRD1 causes drought induction of these genes, which leads to the conclusion that the intact gene works antagonistically with drought in regulating the expression of the set 3 genes. Interestingly, this set of genes is enriched with genes involved in JA response and signaling. Also, we identified that a motif similar to the prolamin-box (Wu et al, 2000) is associated with the set 3 genes and is connected to the JA-related genes, an interaction thus far unknown.

Genes in set 5 are enriched in genes involved in mRNA processing including RNA splicing. These are genes are up-regulated only in the WT but not in *grd1*, indicating that GRD1 potentially mediates drought-regulation of these genes. And, again, we have identified a CRE – an ABRE-like motif – to be associated with these genes, confirming the involvement of RNA spicing in ABA and drought signaling (Papp et al, 2004).

3.4. Discussion

Regulation of gene expression is a very intricate level of cellular control in complex organisms, especially in higher eukaryotes (Komili & Silver, 2007) and the regulatory logic that drives gene expression changes in stress response is governed by the combination of signaling regulators, transcription factors (TFs), cis-regulatory elements or motifs (CREs) and other regulatory molecules (e.g. chromatin modifiers and small RNAs) (Krishnan et al, 2009). Understanding this layer of regulation operating under environmental stress is critical in putting together a roadmap of cellular changes involved in acclimation and resistance. In the study presented here, we have charted out several findings about transcriptional regulation of gene expression under drought stress using a compendium of expression profiles under various drought treatments in Arabidopsis.

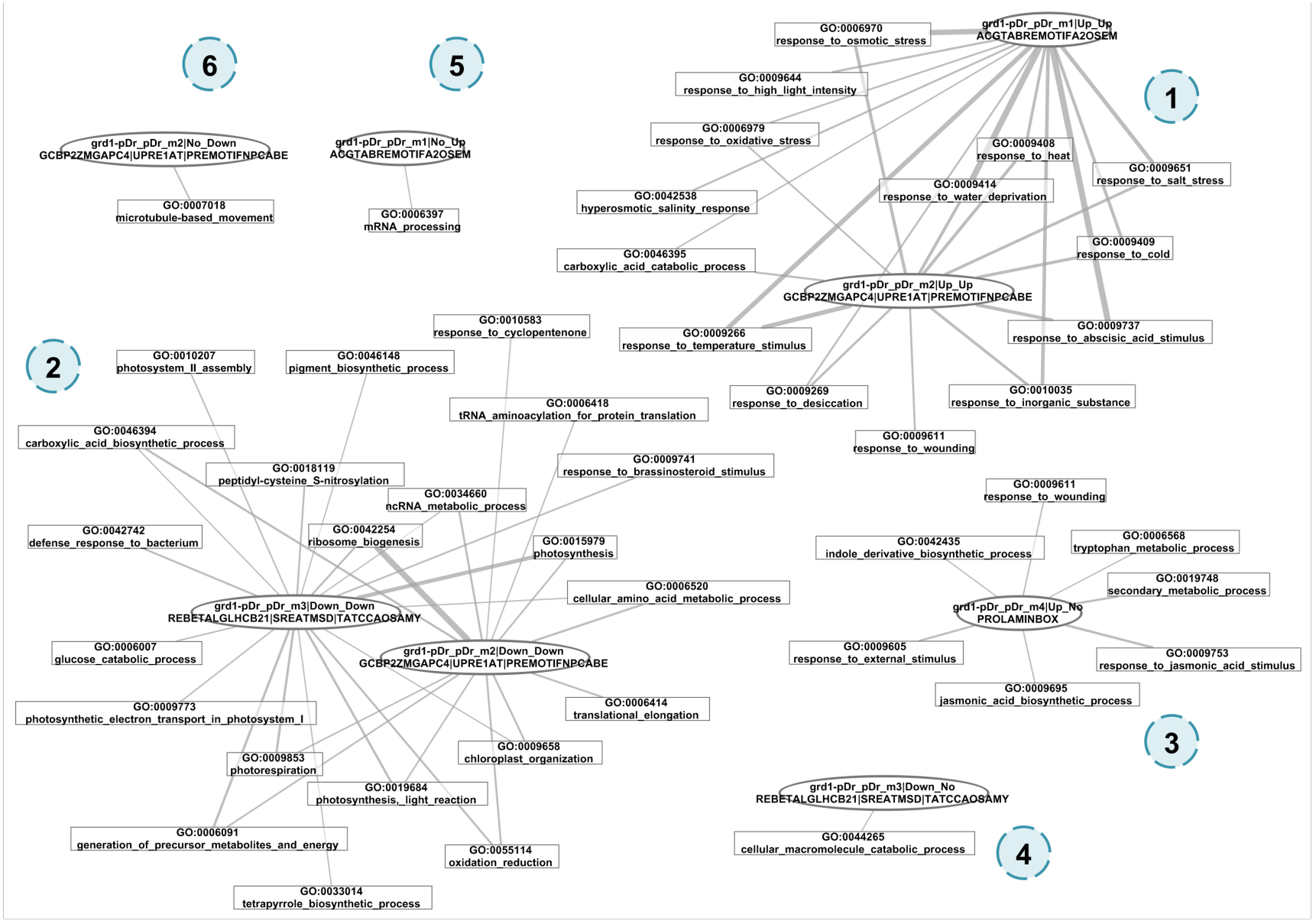

6
5
1
2
3
4
grd1-pDr_pDr_m2|No_Down
GCBP2ZMGAPC4|UPRE1AT|PREMOTIFNPCABE
GO:0007018
microtubule-based_movement
grd1-pDr_pDr_m1|No_Up
ACGTABREMOTIFA2OSEM
GO:0006397
mRNA_processing
GO:0006970
response_to_osmotic_stress
grd1-pDr_pDr_m1|Up_Up
ACGTABREMOTIFA2OSEM
GO:0009644
response_to_high_light_intensity
GO:0006979
response_to_oxidative_stress
GO:0009408
response_to_heat
GO:0009651
response_to_salt_stress
GO:0009414
response_to_water_deprivation
GO:0042538
hyperosmotic_salinity_response
GO:0009409
response_to_cold
GO:0046395
carboxylic_acid_catabolic_process
grd1-pDr_pDr_m2|Up_Up
GCBP2ZMGAPC4|UPRE1AT|PREMOTIFNPCABE
GO:0010583
response_to_cyclopentenone
GO:0009266
response_to_temperature_stimulus
GO:0009737
response_to_abscisic_acid_stimulus
GO:0010207
photosystem_II_assembly
GO:0046148
pigment_biosynthetic_process
GO:0009269
response_to_desiccation
GO:0010035
response_to_inorganic_substance
GO:0006418
tRNA_aminoacylation_for_protein_translation
GO:0009611
response_to_wounding
GO:0046394
carboxylic_acid_biosynthetic_process
GO:0018119
peptidyl-cysteine_S-nitrosylation
GO:0009741
response_to_brassinosteroid_stimulus
GO:0034660
ncRNA_metabolic_process
GO:0009611
response_to_wounding
GO:0042742
defense_response_to_bacterium
GO:0042254
ribosome_biogenesis
GO:0015979
photosynthesis
GO:0042435
indole_derivative_biosynthetic_process
GO:0006568
tryptophan_metabolic_process
GO:0019748
secondary_metabolic_process
GO:0006520
cellular_amino_acid_metabolic_process
grd1-pDr_pDr_m3|Down_Down
REBETALGLHCB21|SREATMSD|TATCCAOSAMY
grd1-pDr_pDr_m4|Up_No
PROLAMINBOX
GO:0006007
glucose_catabolic_process
grd1-pDr_pDr_m2|Down_Down
GCBP2ZMGAPC4|UPRE1AT|PREMOTIFNPCABE
GO:0009605
response_to_external_stimulus
GO:0009753
response_to_jasmonic_acid_stimulus
GO:0009695
jasmonic_acid_biosynthetic_process
GO:0009773
photosynthetic_electron_transport_in_photosystem_I
GO:0006414
translational_elongation
GO:0009658
chloroplast_organization
GO:0009853
photorespiration
GO:0019684
photosynthesis,_light_reaction
grd1-pDr_pDr_m3|Down_No
REBETALGLHCB21|SREATMSD|TATCCAOSAMY
GO:0006091
generation_of_precursor_metabolites_and_energy
GO:0055114
oxidation_reduction
GO:0044265
cellular_macromolecule_catabolic_process
GO:0033014
tetrapyrrole_biosynthetic_process

Figure 3.8: Transcriptional regulatory program underlying the gene expression under drought in the *grd1* mutant compared to WT.
The whole of set of genes regulated under drought in both WT and *grd1* mutant were split into six groups of genes with different patterns of expression as represented in Figure 6. These six genesets were analyzed for enrichment of GO BP terms and informative cis-regulatory motifs using *de novo* analysis. Then, in the context of each of the six groups, associations between motifs and GO BP terms were identified and are presented in the form of a graph. Format is similar to Figure 4. The groups are labeled to match with those in Figure 6. The CREs discovered here are detailed in Fig. 3.S5.

We performed in-depth analysis of genome-wide gene expression under moderate (mDr) and progressive drought (pDr) (Harb et al., 2010, *in press*) and compared it with time course gene expression profiles under acute drought (aDr) (Kilian et al, 2007). At the gene-level we found that there is minimal overlap between mDr and aDr, while pDr shows a reasonable amount of overlap with aDr response, especially at later time points (Fig. 3.1). But, together, there is enormous diversity in drought-response with 11, 660 genes (~43% of the total genes in the genome) perturbed by one form of drought or another. However, apart from large difference between the treatments itself, difference in the growth conditions of the plants, the developmental stage (Catala et al, 2007) and time of day at sampling (Wilkins et al, 2010), and the tissue sampled (Dinneny et al, 2008) between our (mDr and pDr) experiments and AtGenExpress aDr will contribute to the observed differences in gene expression under stress. In spite of these factors potentially confounding our results, several insights about gene expression under drought stress have emerged due to an integrative analysis carried out here.

First, we have identified several biological processes that are perturbed similarly and specifically across drought treatments. Drought affected processes cut across all levels of cellular organization and complexity including hormonal signaling, DNA packaging, transcription, mRNA processing, rRNA processing, ribosome biogenesis, tRNA processing, amino acid biosynthesis, translational initiation and elongation, protein folding and localization, energy production and storage, secondary metabolism, structural protein organization, immune response and developmental pathways.

Second, to shed light on the regulatory network that underlies the regulation of these biological processes, we used *de novo* cis-regulatory element (CRE) discovery followed by mining of transcriptional regulatory programs (Krishnan & Pereira, 2008). CREs mediating the interaction of TFs with their target genes are sites of high rates of evolutionary diversity driving the

progressively complex gene expression correlating with organismal complexity (Wray, 2007). Hence, without restricting ourselves to known CREs in the their forms curated in databases, we used the *de novo* approach to discover CREs only using genome sequence information and gene expression patterns. This analysis led to the identification and refinement of several known CREs in the context of drought and discovery of novel ones.

The ABRE-like motifs, known target-binding sites for the basic leucine zipper (bZIP) TFs (Shinozaki & Yamaguchi-Shinozaki, 2000) are among the most prominent motifs associated with the core set of stress response genes up-regulated by most drought conditions. There is general acceptance that a single ACGT-containing ABRE is insufficient to facilitate ABA-induced expression (Shen et al, 2004), and these motifs were indeed predicted to mediate the regulation of several different groups of processes in combination with other motifs like DRE and pyrimidine-box. This is in agreement with several examples of the combinatorial control involved in ABA response (Abe et al, 2003; Chung & Parish, 2008; Narusaka et al, 2003; Simpson et al, 2003).

The coupling element 3 (CE3) – ACGCGTGTCCTC – identified among the down-regulated genes in mDr-Day10 and aDr-15min is very similar to ABRE and is again part of the CRE module (ABA response complex) that is necessary and sufficient for ABA-induced transcription (Shen et al, 2004). CE3 is thought to be functionally equivalent to ABRE (Hobo et al, 1999), but, based on our results, we hypothesize that the functional equivalence with ABRE need not always hold and there might be conditions (like mDr-Day1 or aDr-25min) where this element could mediate down-regulation of genes.

The pyrimidine-box-like motifs identified to be involved in combinatorial regulation with ABRE among are typically found among gibberellic acid (GA) regulated genes and is a part of the a group of three cis-elements mediating GA-responsive gene expression (tripartite GA-responsive complex; GARC) (Mena et al, 2002; Rubio-Somoza et al, 2006). However, as stated earlier, given that GA and ABA have antagonistic roles, it is unclear how this motif will cooperate with ABRE to regulate stress gene expression. Interestingly, the minus314-like motif predicted to be

involved with ABRE in gene up-regulation in aDr-30min-1h is also very similar to the pyrimidine box.

The unfolded-protein response has been implicated in drought or other abiotic stress response in several studies (Alvim et al, 2001; Irsigler et al, 2007; Liu et al, 2007). In line with this knowledge, we have identified UPRE-related motifs among the mDr-Day1 up-regulated genes and among pDr-regulated genes. Since the transcriptional up-regulation of a large number of genes involved in the UPR pathway is the most prominent strategy for the cell to cope with ER stress (Urade, 2007), we hypothesize that the repression of 'protein folding' observed in the different drought treatments triggers the expression of UPR genes via the identified UPRE motif.

Among the down-regulated programs, the Ibox motif was predicted to mediate the down-regulation of several processes including chlorophyll biosynthesis and photosynthesis, which is expected based on the fact that the Ibox is enriched among the light-responsive genes (Giuliano et al, 1988). Ribosome biogenesis and related processes that are consistently repressed by different drought treatments were predicted to be regulated by TFs that bind to site II and Telo-box motifs. The site II motifs are recognized by TFs of the TCP family and have been confirmed to be important in the regulation of ribosome protein (RP) genes in combination with the telo-box motif (Tremousaygue et al, 2003). These motifs are co-located in the promoters of about 70% of 216 ribosomal protein genes in Arabidopsis. In addition, there is evidence that the site II motifs also possibly coordinate the expression of nuclear genes encoding components of the mitochondrial oxidative phosphorylation machinery in both Arabidopsis and rice (Welchen & Gonzalez, 2006). Therefore, this program involving site II and telo-box motifs could mediate the down-regulation of major processes that affect protein production under drought stress.

Using a similar approach, we have dissected drought response in the *grd1* mutant (knockout of GRD1 gene that encode a MYB TF, which is induced by drought) to identify processes that are differently regulated in the mutant compared to WT. mRNA processing including RNA splicing was observed to be up-regulated only in the WT but not in *grd1*, indicating that GRD1 potentially mediates drought-induction of these genes. An ABRE-like motif was also associated with this set of genes giving credence to the regulation of these genes under drought. RNA

splicing has been implicated yet another level of regulation in abiotic stress response (Mazzucotelli et al, 2008) and stress signaling has been found involve changing alternative splicing profiles (Iida et al, 2004; Reddy, 2007). Furthermore, ABA signaling and response itself is dependent on a sound RNA processing and splicing machinery (Hirayama & Shinozaki, 2007; Papp et al, 2004; Tillemans et al, 2006). Thus, given the importance of RNA splicing in drought response and that *grd1* is drought sensitive, we hypothesize that GRD1 is a regulator of proper RNA splicing and processing that aids to fine-tune plant responses to drought stress.

Another interesting process differently regulated in *grd1* is the up-regulation of JA-response genes only in *grd1* and not in the WT. This observation means that in the absence of GRD1, drought causes induction of these genes, suggesting that the intact gene works antagonistically with drought in regulating the expression of JA-response genes. JA is induced under drought and is generally thought to lead to negative effects involving repression of photosynthesis and cell growth (Wasternack, 2007). Based on these observations, we propose that GRD1 might act as a negative regulator of JA biosynthesis and signaling under drought stress to improve plant response and resistance to drought.

Stress-responsive transcriptional responses are specific to a metabolic or developmental state and activate stress-responsive mechanisms that exist in the cell. In addition to these 'hard-wired' responses, transcriptional programs show considerable plasticity to adapt to a wide range of stresses, including those not encountered during evolutionary history (Lopez-Maury et al, 2008). Using a highly integrative approach making minimal assumptions and taking complete advantage of existing data, we have described several transcriptional regulatory programs that underlie drought stress response in Arabidopsis and proposed specific regulatory mechanisms that can be experimentally verified.

3.5. Methods

3.5.1. Expression Profiling of Early and Late mDr and pDr

Briefly, for mDr, plants at 30 days after sowing were subjected to a moderate drought treatment and RNA was isolated from two biological samples of 5 pooled plants each were collected at days 1 and 10. For pDr, five weeks WT plants were drought-treated until the first day of wilting

and RNA was isolated from drought-treated and control plants of each genotype (WT and *grd1*). Two replications per treatment for mDr and pDr were used for Affymetrix (ATH1 25K) GeneChip hybridization. aDr shoot expression data was obtained from NCBI GEO (Barrett et al, 2009) accessions GSE5624 (drought) and GSE5620 (control).

3.5.2. Analysis of Gene Expression Profiles

For each of the drought experiments, mDr-Day1, mDr-Day10, pDr, *grd1*-Dr and aDr (seven time points) raw data were background corrected, normalized and summarized according to the custom CDF (see below) using RMA (Irizarry et al., 2003; Ihaka and Gentleman, 1996; Gentleman et al., 2004), followed by non-specific filtering of genes that do not have enough variation (interquartile range (IQR) across samples < IQR_{median}) to allow reliable detection of differential expression. A linear model was then used to detect differential expression of the remaining genes on comparing drought treated samples versus well-watered control (Smyth 2004). The *p*-values from the moderated *t*-tests were converted to *q*-values to correct for multiple hypothesis testing (Storey and Tibshirani, 2003), and genes with *q*-value <0.1 were considered as differentially expressed in response to the mDr and pDr drought treatments. This *q*-value cut-off was used to ensure that there were enough number of genes in mDr Day10 to analyze further based on functional enrichment. Genes with q-value <0.05 were chosen for aDr. This cut-off this based on previous studies using this data.

3.5.3. Reannotation of Arabidopsis ATH1 Probe-Gene Mapping

The mapping of Affymetrix ATH1 probe sets to Arabidopsis loci provided TAIR is arrived at using the following procedure (ftp://ftp.arabidopsis.org/Microarrays/Affymetrix/README on 7/30/09): The mapping to the TAIR8 Transcripts was performed using the BLASTN program with E-value cutoff <= 9.9e-6. For the 25-mer oligo probes used on the Affymetrix chips, the required match length to achieve this E-value is 23 or more identical nucleotides. To assign a probe set to a given locus, at least 9 of the probes included in the probe set were required to match a transcript at that locus. Disregarding probe sets that map to more than one locus, this procedure results in mapping 21,180 probe sets to 21,019 genes.

TAIR as a database will wish to preserve the probe-to-'probe set' definitions provided by Affymetrix for users to map probe sets to genes after performing microarray analysis using the default chip definition file (CDF). But strictly, there are two issues in this procedure that could lead to significant inaccuracies in estimation of gene expression: a) A probe set mapped to a locus can contain up to 2 probes that do not match the locus at all and other probes that do not match the locus uniquely; b) Since multiple probe sets can map to the same locus, during analysis one has to make an ad hoc procedure to either combine information from all the mapping probe sets or choose one of the probe sets based on an arbitrary criterion. Both choices have been used in previous studies frequently.

To get around these issues and improve to the mapping generally, we sought to: a) Increase the stringency of mapping a 25-mer probe from 23 or more identical nucleotides to a *perfect* match; b) Assign a probe to a locus only when it uniquely maps to that locus; c) Combine all the probes that uniquely map to a given locus into a single probe set, identified after the locus.

Thus, a high-quality custom CDF was built for the Arabidopsis ATH1 array by uniquely mapping 232,697 probe sequences (http://www.affymetrix.com/analysis/downloads_/data/) to 21,389 Arabidopsis (TAIR8) gene-based probe sets in the following manner: (i) probes that have perfect sequence identity with a single target gene were selected, (ii) probes mapping to reverse complements of genes were annotated separately as antisense probes (not used in the above counts), and finally, (iii) probes were grouped into probe sets, each corresponding to a single gene, and probe sets having at least 3 unique sequences were retained (>99% probe sets have >=5 probes). Note that these stringent criteria used to construct the CDF make it possible to reliably measure expression values of members of multigene families (free from cross-hybridization between paralogs showing high sequence similarity).

3.5.4. Promoter Analysis

For analysis of potential promoter-resident cis-regulatory elements (CREs), FIRE (Elemento et al, 2007) was used to discover motifs informative about the different sets of differentially expressed genes compared to the rest of the genes in the genome. Briefly, FIRE seeks to discover motifs whose patterns of presence/absence across all considered regulatory regions (motif

profile) are most informative about the expression of the corresponding genes (expression profile). To measure these associations, FIRE uses mutual information (MI) (Cover & Thomas, 2006). FIRE performs a randomization test and considers an observed MI value (for a motif-expression profile pair) to be significant only when it is greater than all the random MI values calculated by randomly assigning the expression values to genes. A Z-score reflecting how far the observed MI value is, in number of standard deviations, from the average random MI is calculated. These are the Z-scores presented for each motif in Figure 3.3. Moreover, it also performs jack-knife resampling (Efron, 1979), where, in each of 10 trial the above randomization test is carried out. Only motifs that are statistically significant in at least 6 trials are reported. Newly discovered motifs were compared to known cis-elements in the PLACE database (Higo et al, 1999) and to each other using STAMP (Mahony & Benos, 2007). All upstream sequences were obtained from TAIR. This *de-novo* approach was taken since i) CREs could diverge far more quickly than coding sequences across species, making them hard to find simply by searching, and ii) searching based on known elements in Arabidopsis is limited by the scope of experimental identification in a select set of genes, making identification of degenerate yet potentially functional positions in the element hard.

3.5.5. Discovery of transcriptional regulatory programs

Gene function descriptions and GO biological process (BP) annotations were downloaded from TAIR (TAIR8; Swarbreck et al., 2008). Applying the true-path-rule, a gene annotated with a particular GO term was also annotated with all its ancestors. To avoid very generic, non-informative terms for further analysis, only terms annotating <=500 genes ('specific GO-terms') were retained. Genes annotated with a given specific GO-term were considered as a gene set. Genes containing CREs discovered *de novo* were included as additional gene sets. Drought regulated genesets were defined as sets of genes up- and down-regulated in mDr-Day1, mDr-Day10 and pDr. From the seven time points for aDr – 0.25h, 0.5h, 1h, 3h, 6h, 12h and 24h – based on gene expression similarity (data not shown) aDr_0.5h and aDr_1h, and aDr_3h and aDr_6h were combined into aDr_0.5-1h and aDr_3-6h, respectively, and five aDr genesets were obtained.

The GO_BP and CRE genesets described above were tested for the statistical significance of enrichment among the experimentally identified drought gene sets (mDr-Day1, mDr-Day10, pDr, aDr_0.25h, aDr_0.5-1h, aDr_3-6h, aDr_12h and aDr_24h up- and down-regulated genes) using the cumulative hypergeometric test. For a pair of gene sets *i* and *j*, if *N* is the total number of genes, n_i and n_j are the number of genes in gene set *i* and *j*, and *m* is the number of genes common to the gene sets, the probability (*p*-value) of an overlap (enrichment) of size equal to or greater than observed is given by the formula below.

$$P(X = x \geq m) = \sum_{x=m}^{\min(n_i,n_j)} \frac{\binom{n_i}{x}\binom{N-n_i}{n_j-x}}{\binom{N}{n_j}}$$

To adjust for multiple comparisons, a Benjamini-Hochberg false discovery rate (FDR; *q*-value) (Benjamini & Hochberg, 1995) was calculated from the *p*-values, and a *q*-value threshold of 0.01 was used for significance.

To find associations between GO_BP and CREs in the context of a drought geneset (say 'pDr_Up'), instead of the all the genes in the genome, the universe of genes was restricted to that drought geneset and overlap between the GO_BP genesets and CRE-genesets within this universe was evaluated using the cumulative hypergeometric test. For mDr, pDr and aDr analyses, GO_BP-CRE pairs with q-value <0.01 were considered significant. For *grd1* analysis, since we were interested in deciphering some link between a CRE and GO_BP term in groups 3, 4, 5 and 6 (Figure 3.6) a more relaxed *q*-value cutoff (*q*-value <0.1) was used to discover GO_BP-CRE pairs. All these associations were represented as a graph where GO_BP terms are connected to their enriched CRE partner using Cytoscape (Shannon et al, 2003).

3.6. References

Abe H, Urao T, Ito T, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* **15:** 63-78

Alvim FC, Carolino SM, Cascardo JC, Nunes CC, Martinez CA, Otoni WC, Fontes EP (2001) Enhanced accumulation of BiP in transgenic plants confers tolerance to water stress. *Plant Physiol* **126:** 1042-1054

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37:** D885-890

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* **57:** 289-300

Boyer JS (1982) Plant productivity and environment. *Science* **218:** 443-448

Catala R, Ouyang J, Abreu IA, Hu Y, Seo H, Zhang X, Chua NH (2007) The Arabidopsis E3 SUMO ligase SIZ1 regulates plant growth and drought responses. *Plant Cell* **19:** 2952-2966

Chung S, Parish RW (2008) Combinatorial interactions of multiple cis-elements regulating the induction of the Arabidopsis XERO2 dehydrin gene by abscisic acid and cold. *Plant J* **54:** 15-29

Cover TM, Thomas JA (2006) *Elements of information theory*, 2nd edn. Hoboken, N.J.: Wiley-Interscience.

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* **4:** 25

Dinneny JR, Long TA, Wang JY, Jung JW, Mace D, Pointer S, Barron C, Brady SM, Schiefelbein J, Benfey PN (2008) Cell identity mediates the response of Arabidopsis roots to abiotic stress. *Science (New York, NY)* **320:** 942-945

Efron BL (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7:** 1-26

Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* **28:** 337-350

Giuliano G, Pichersky E, Malik VS, Timko MP, Scolnik PA, Cashmore AR (1988) An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. *Proc Natl Acad Sci U S A* **85:** 7089-7093

Hannah MA, Heyer AG, Hincha DK (2005) A global survey of gene regulation during cold acclimation in Arabidopsis thaliana. *PLoS Genet* **1:** e26

Hattori T, Totsuka M, Hobo T, Kagaya Y, Yamamoto-Toyoda A (2002) Experimentally determined sequence requirement of ACGT-containing abscisic acid response element. *Plant Cell Physiol* **43:** 136-140

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

Hirayama T, Shinozaki K (2007) Perception and transduction of abscisic acid signals: keys to the function of the versatile plant hormone ABA. *Trends Plant Sci* **12:** 343-351

Hobo T, Asada M, Kowyama Y, Hattori T (1999) ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. *Plant J* **19:** 679-689

Huang D, Wu W, Abrams SR, Cutler AJ (2008) The relationship of drought-related gene expression in Arabidopsis thaliana to hormonal and environmental factors. *Journal of Experimental Botany* **59:** 2991-3007

Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K (2004) Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. *Nucleic Acids Res* **32:** 5096-5103

Irsigler AS, Costa MD, Zhang P, Reis PA, Dewey RE, Boston RS, Fontes EP (2007) Expression profiling on soybean leaves reveals integration of ER- and osmotic-stress pathways. *BMC Genomics* **8:** 431

Jiao Y, Lau OS, Deng XW (2007) Light-regulated transcriptional networks in higher plants. *Nat Rev Genet* **8:** 217-230

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

Kim KN, Guiltinan MJ (1999) Identification of cis-acting elements important for expression of the starch-branching enzyme I gene in maize endosperm. *Plant Physiol* **121:** 225-236

Komili S, Silver PA (2007) Coupling and coordination in gene expression processes: a systems biology view. *Nat Rev Genet*

Kreps JA, Wu Y, Chang HS, Zhu T, Wang X, Harper JF (2002) Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress. *Plant Physiol* **130:** 2129-2141

Krishnan A, Ambavaram MMR, Harb A, Batlang U, Wittich PE, Pereira A (2009) Genetic networks underlying plant abiotic stress responses. In *Genes for Plant Abiotic Stress*, Jenks MA, Wood AJ (eds). John Wiley & Sons, Inc., Ames IA, USA.

Krishnan A, Pereira A (2008) Integrative approaches for mining transcriptional regulatory programs in Arabidopsis. *Brief Funct Genomic Proteomic* **7:** 264-274

Liu JX, Srivastava R, Che P, Howell SH (2007) Salt stress responses in Arabidopsis utilize a signal transduction pathway related to endoplasmic reticulum stress signaling. *Plant J* **51:** 897-909

Lopez-Maury L, Marguerat S, Bahler J (2008) Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat Rev Genet* **9:** 583-593

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

Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* **35:** W253-258

Martinez IM, Chrispeels MJ (2003) Genomic analysis of the unfolded protein response in Arabidopsis shows its connection to important cellular processes. *Plant Cell* **15:** 561-576

Mazzucotelli E, Mastrangelo AA, Crosatti C, Guerra D, Stanca AM, Cattivelli L (2008) Abiotic stress response in plants: When post-transcriptional and post-translational regulations control transcription. *Plant Sci* **174:** 420-431

Mena M, Cejudo FJ, Isabel-Lamoneda I, Carbonero P (2002) A role for the DOF transcription factor BPBF in the regulation of gibberellin-responsive genes in barley aleurone. *Plant Physiol* **130:** 111-119

Narusaka Y, Nakashima K, Shinwari ZK, Sakuma Y, Furihata T, Abe H, Narusaka M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. *Plant J* **34:** 137-148

Papp I, Mur LA, Dalmadi A, Dulai S, Koncz C (2004) A mutation in the Cap Binding Protein 20 gene confers drought tolerance to Arabidopsis. *Plant Mol Biol* **55:** 679-686

Piskurewicz U, Tureckova V, Lacombe E, Lopez-Molina L (2009) Far-red light inhibits germination through DELLA-dependent stimulation of ABA synthesis and ABI3 activity. *EMBO J* **28:** 2259-2271

Reddy AS (2007) Alternative splicing of pre-messenger RNAs in plants in the genomic era. *Annu Rev Plant Biol* **58:** 267-294

Rubio-Somoza I, Martinez M, Diaz I, Carbonero P (2006) HvMCB1, a R1MYB transcription factor from barley with antagonistic regulatory functions during seed development and germination. *Plant J* **45:** 17-30

Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T, Satou M, Akiyama K, Taji T, Yamaguchi-Shinozaki K, Carninci P, Kawai J, Hayashizaki Y, Shinozaki K (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant J* **31:** 279-292

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498-2504

Shen QJ, Casaretto JA, Zhang P, Ho TH (2004) Functional definition of ABA-response complexes: the promoter units necessary and sufficient for ABA induction of gene expression in barley (Hordeum vulgare L.). *Plant Mol Biol* **54:** 111-124

Shinozaki K, Yamaguchi-Shinozaki K (2000) Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Curr Opin Plant Biol* **3:** 217-223

Shinozaki K, Yamaguchi-Shinozaki K (2007) Gene networks involved in drought stress response and tolerance. *J Exp Bot* **58:** 221-227

Shinozaki K, Yamaguchi-Shinozaki K, Seki M (2003) Regulatory network of gene expression in the drought and cold stress responses. *Curr Opin Plant Biol* **6:** 410-417

Simpson SD, Nakashima K, Narusaka Y, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Two different novel cis-acting elements of erd1, a clpA homologous Arabidopsis gene function in induction by dehydration stress and dark-induced senescence. *Plant J* **33:** 259-270

Stintzi A, Weber H, Reymond P, Browse J, Farmer EE (2001) Plant defense in the absence of jasmonic acid: the role of cyclopentenones. *Proc Natl Acad Sci U S A* **98:** 12837-12842

Supper J, Strauch M, Wanke D, Harter K, Zell A (2007) EDISA: extracting biclusters from multiple time-series of gene expression profiles. *BMC Bioinformatics* **8:** 334-334

Tillemans V, Leponce I, Rausin G, Dispa L, Motte P (2006) Insights into nuclear organization in plants as revealed by the dynamic distribution of Arabidopsis SR splicing factors. *Plant Cell* **18:** 3218-3234

Tremousaygue D, Garnier L, Bardet C, Dabos P, Herve C, Lescure B (2003) Internal telomeric repeats and 'TCP domain' protein-binding sites co-operate to regulate gene expression in Arabidopsis thaliana cycling cells. *Plant J* **33:** 957-966

Urade R (2007) Cellular response to unfolded proteins in the endoplasmic reticulum of plants. *FEBS J* **274:** 1152-1171

Walther D, Brunnemann R, Selbig J (2007) The regulatory code for transcriptional response diversity and its relation to genome structural properties in A. thaliana. *PLoS Genetics* **3:** e11-e11

Wasternack C (2007) Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Ann Bot* **100:** 681-697

Welchen E, Gonzalez DH (2006) Overrepresentation of elements recognized by TCP-domain transcription factors in the upstream regions of nuclear genes encoding components of the mitochondrial oxidative phosphorylation Machinery. *Plant Physiol* **141:** 540-545

Wilkins O, Brautigam K, Campbell MM (2010) Time of day shapes Arabidopsis drought transcriptomes. *Plant J*

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8:** 206-216

Wu C, Washida H, Onodera Y, Harada K, Takaiwa F (2000) Quantitative nature of the Prolamin-box, ACGT and AACA motifs in a rice glutelin gene promoter: minimal cis-element requirements for endosperm-specific gene expression. *Plant J* **23:** 415-421

Yamaguchi-Shinozaki K, Shinozaki K (2005) Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends Plant Sci* **10:** 88-94

4. The state of network-based gene function prediction in Arabidopsis

Arjun Krishnan, T. M. Murali, Brett Tyler, Andy Pereira

4.1. Abstract

Arabidopsis is a primary model organism for plants. Understanding its cellular machinery and transferring the resulting knowledge to plants of economic importance is a key challenge in biology. However, poor knowledge of the cellular roles played by Arabidopsis genes remains a fundamental barrier to achieving this goal. Currently, of the approximately 29,000 genes in Arabidopsis, only about 50% are annotated with any function (Gene Ontology biological process), with annotations for less than one-third of these genes coming from experimentally verified sources. Networks of functional linkages between genes and gene products, built by integrating high-throughput functional genomic data, are an attractive approach for alleviating this problem. These networks can be used in conjunction with existing functional annotations as a powerful tool for predicting functions of currently un-annotated or poorly annotated genes. In this study, we examine the state of network-based gene function prediction in Arabidopsis by evaluating the performance of various algorithms on *AraNet*, a recently developed gene interaction network for Arabidopsis. These algorithms take advantage of local interactions or overall network structure to transfer and predict gene function. We first explore the influence of the number of genes annotated to a function and the source of annotation evidence (experimental or electronic) on prediction performance. Upon identifying the candidate best algorithm, we find topological properties of genes in the network that correlate well with the performance of the algorithm. We then measure the ability of the network-based approach to predict conserved and plant-specific functions. We discuss specific examples of both conserved and plant-specific functions that are most and least 'predictable'. Finally, we provide guidelines to the plant community on the development of more refined functional linkage networks, improved methods for computational prediction of gene function, and strategies for prioritization of experiments to verify predictions.

4.2. Introduction

Arabidopsis has been the primary model organism for the plant kingdom for twenty-five years (Koornneef & Meinke, 2010). It was the third eukaryotic genome to be sequenced. The Arabidopsis genome harbors about 29, 000 genes (similar to the number of genes in the human genome). Discovering the function of each of these genes is critical for understanding the overall organization of the cellular machinery that supports complex processes in plants such as growth and development, energy production, and response to environment. The *function* of a gene has multiple definitions depending on the context (Bork et al, 1998), spanning concepts including the developmental or growth stage at which the gene is expressed, the molecular function of the gene's protein product, the signaling/metabolic/regulatory pathway the protein might participate in, the nature of the gene's response to the environment, or an observable phenotypic trait it is associated with. Gene Ontology (GO) annotations (Ashburner et al, 2000), especially those in the biological process (BP) hierarchy, cut across many of these levels. Hence, in this work, we use GO BP annotations as our basis for defining gene function.

Although numerous experiments have been done in Arabidopsis to elucidate gene functions, only about 15% of the genes in the genome have any experimentally verified functional annotation (Fig. 4.1). An additional 20% of the genes have annotations based on author statements, expert curation, or sequence/structural similarity to genes with BP annotations in other organisms. Approximately 14% of the genes have 'electronic' annotations not assigned by a curator, leaving more than 50% of the genes in the genome not annotated with any function. When annotations to only 'specific' functions (defined in this work as GO BP terms annotating <300 genes, which is approximately 1% of the number of genes in the Arabidopsis genome) are considered, the percentage of genes annotated with any function drops to 26. This lack of specific knowledge about the functions of nearly 74% of Arabidopsis genes poses a great challenge to systems biology studies of this model plant.

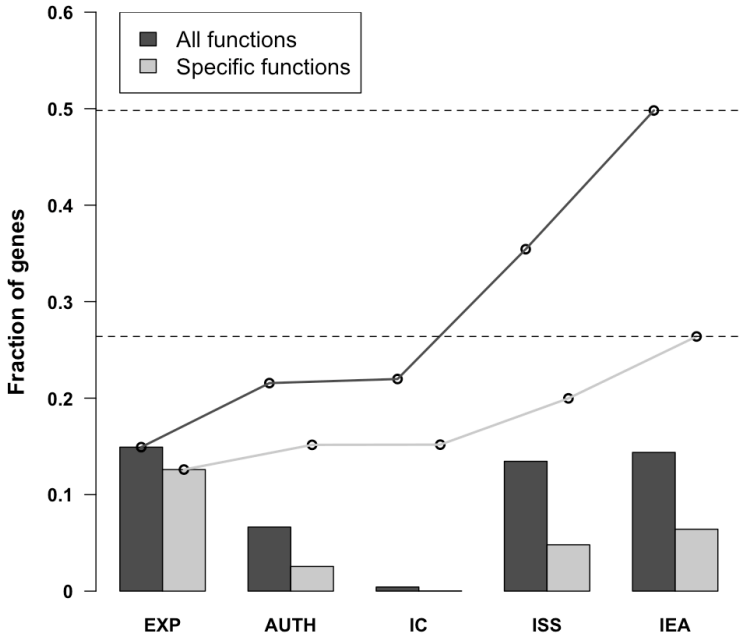


Figure 4.1: Represent the fraction of genes (out of a total of 29, 889 genes in the genome) annotated with all (dark grey) or 'specific' (light grey) functions based upon different sources of evidence (GO evidence codes).
Specific functions are defined as GO BP terms that annotate <300 genes (approximately 1% of the genes in the genome). Evidence codes are grouped as follows: EXP: Experimental evidence codes IDA, IMP, IPI, IGI and IEP; AUTH: Author statement evidence codes TAS and NAS; IC: Inferred by Curator; ISS: Inferred from Sequence or Structural Similarity; and IEA: Inferred from Electronic Annotation. See http://www.geneontology.org/GO.evidence.shtml for details on these codes. Evidence codes are ordered from left to right in decreasing relation to experimentally validated annotations. The line graphs represent the cumulative fractions, culminating in the maximum denoted by the dashed lines. Genes counted in an evidence code group are not included in any later group. Counts are based on GO BP annotations obtained after applying the true-path-rule.

Molecular interaction networks are powerful frameworks for representing the ways in which genes, proteins, metabolites and other molecules work together to perform specific functions inside the cell (Yamada & Bork, 2009). Typically, interactions in these networks are "functional", i.e., two genes are directly connected if there is some evidence that they may participate in the same biological process. Functional interaction networks have been derived from an amalgam of diverse datasets in multiple unicellular and multi-cellular organisms (Costello et al, 2009; Guan et al, 2008; Hu et al, 2009; Huttenhower et al, 2009; Lee et al, 2008; Zhu et al, 2008). In such networks, the interactions that a gene participates in establish the putative functional context within which that gene operates in the cell. Therefore, when functions for some genes in this network have been identified through rigorous experimentation, other less-studied genes can also be associated with the function of the genes they interact with, based on the principle of guilt-by-association (Sharan et al, 2007). Recently, a genome-scale functional interaction network, termed *AraNet*, was published for Arabidopsis based on the integration of

24 network datasets: five from Arabidopsis and the rest from human, fly, worm and yeast (Lee et al, 2010). The datasets comprise of co-citation, mRNA coexpression, domain co-occurrence, gene-neighborhood, genetic interactions, literature curated interactions, affinity purification/mass spectroscopy, phylogenetic profiles, protein tertiary structure and yeast-two-hybrid assays. The authors use AraNet to prioritize genes for limited-scale functional screening concerning a "trait" (defined using descriptive GO biological process) of interest: necessarily gene-function-prediction. Taking a 'local' network neighborhood approach, for an uncharacterized gene they assign a score for each BP that is equal to the sum of the weights of the interactions linking the gene to other genes already annotated by the term. AraNet represents the largest integration effort so far to construct a functional linkage network for Arabidopsis. Hence, it is a rich resource for functional characterization of the Arabidopsis genome.

In the research presented here, we use AraNet to assess the state of network-based gene function prediction in Arabidopsis. We employ six prediction methods (five distinct algorithms, with one in two flavors) on AraNet and measure their performance in making high-confidence predictions for 374 specific biological processes based on 5-fold cross-validation (Fig. 4.2). Three of the algorithms are variations on the guilt-by-association approach used in the AraNet publication. The remaining three prediction algorithms (SinkSource (introduced here), FunctionalFlow (Nabieva et al, 2005), and Hopfield networks (Karaoz et al, 2004)) operate by applying guilt-by-association-like principles repeatedly over the entire network. These algorithms have the ability to propagate functional annotations across the network in a controlled manner, while taking both short-range and long-range connections within the network into account.

4.3. Results

In this work, we use GO BP annotations as our basis for defining gene function. By design, there are several BP annotation terms that describe very general cellular phenomena (e.g. 'biopolymer modification', 'transport' or 'cell communication') and annotate numerous genes as a result. We reasoned that genes predicted to belong to very specific BPs might be more amenable to experimental verification. Therefore, using the number of annotated genes as a guide to identify such specific functions, we select 374 GO BP terms that annotate <300 genes (approximately 1% of the total genes in the genome) (see *Methods*) for further analysis.

We apply the semi-supervised learning pipeline depicted in Figure 4.2 to assess the performance of network-based gene function prediction algorithms in predicting these specific functions in Arabidopsis based on the AraNet gene functional interaction network (Lee et al, 2010). This network contains 19,647 genes and 1,062,222 edges, with edge weights corresponding to the log likelihood score (LLS) of interaction. We apply this pipeline to annotations based on two sets of GO evidence codes (ECs): one containing gene annotations based on all ECs, and the other excluding annotations based on computational (ISS) or electronic (IEA) ECs. Annotations based on IEA are not manually curated, they may harbor numerous false positive annotations. Although annotations based on ISS are curator-assigned, these annotations still require experimental confirmation and could harbor false positives, although to an extent lesser than annotations based on IEA. Therefore, we define two sets of ECs – 'all ECs' and 'sans ISS and IEA' – to explore the effect of including computational/automated annotations on the quality of gene function predictions.

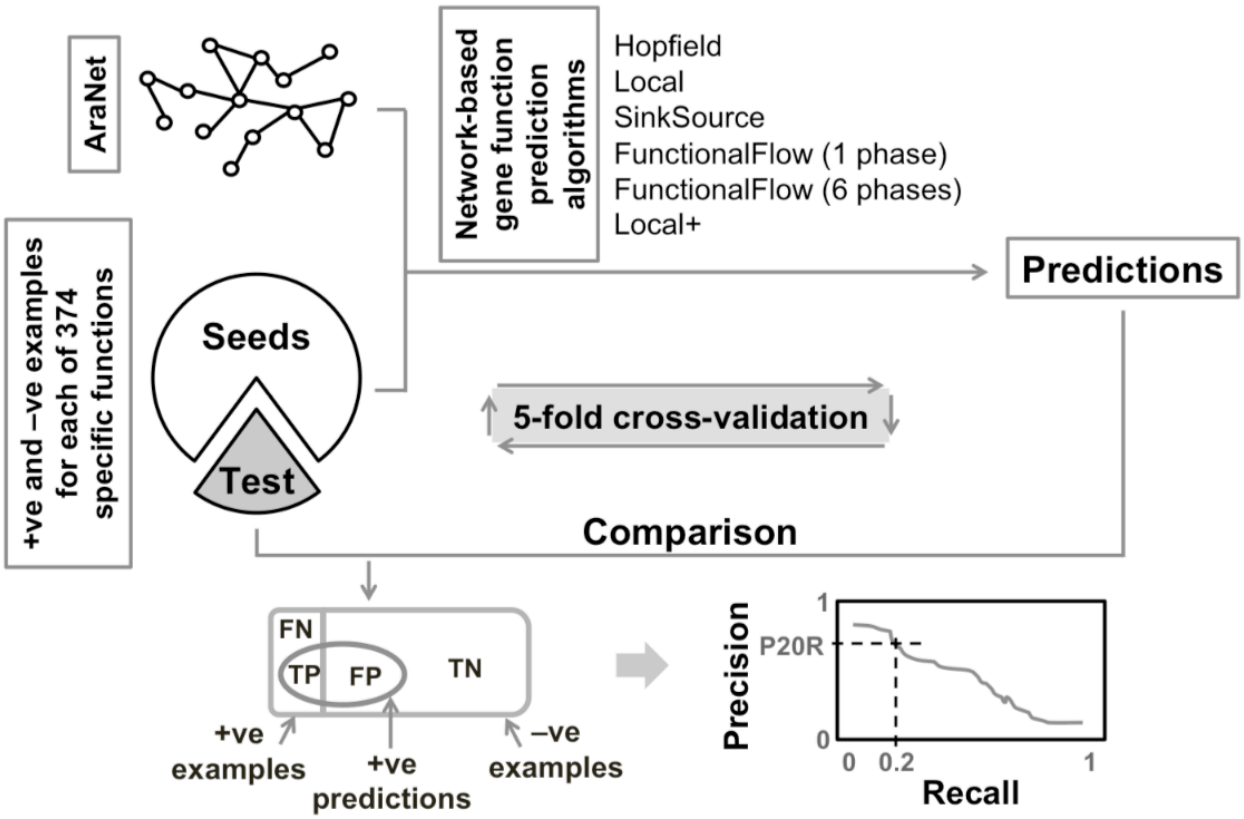


Figure 4.2: Summary of the approach used in this study to assess network-based gene function prediction in Arabidopsis.
We execute a 5-fold cross-validation scheme for each of the 374 specific functions. We run the entire pipeline once using functional annotations based on all evidence codes (all ECs) and once after excluding computational/electronic annotations (Sans ISS IEA). Positive examples (with respect to a function *f*): set of genes annotated with function *f* (after applying the true-path-rule); Negative examples: set of genes not annotated with *f* or any of its ancestors; TP: True positives; FP: False positives; TN: True negatives; FN: False negatives; Precision: Fractions of positive predictions that are correct; Recall: Fraction of known positives that have been predicted correctly; P20R: Precision at 20% recall. See main text for details.

For each function *f*, we define positive examples as the set of genes that are annotated with *f* (after applying the true-path-rule) and negative examples as the set of genes not annotated to *f* or to any of its ancestors. We mark all the other genes as 'unknown' examples for *f*. To predict gene function using these examples in the context of a network, we apply the following algorithms, which we summarize briefly here (see *Methods* for details):

1. SinkSource: The following physical analogy explains this algorithm. We consider the underlying functional interaction network to be a flow network. Here, each edge is a pipe and its weight denotes the amount of fluid that can flow through the pipe per unit time. Each node has a reservoir of fluid. We maintain the level of the reservoir at each positive example at 1 unit and at each negative example at 0 units. We let fluid flow through this network. At equilibrium (when the amount of fluid flowing into each node is equal to the amount flowing out), the reservoir height at each node denotes our confidence that the node is annotated with that function.
2. Hopfield networks (Karaoz et al, 2004): This algorithm is similar to SinkSource, except that the reservoir level at each negative example is set to -1 and the reservoir level at each unknown example is thresholded to 1 or to -1.
3. FunctionalFlow (Nabieva et al, 2005): This algorithm also treats the functional interaction network as flow network. It does not use negative examples and sets the reservoir level at positive examples to be infinity, permitting the reservoir level at an unknown example to increase without bound. Hence, the algorithm stops after a user-specified number of phases.
4. FunctionalFlow run for a single phase is equivalent to the guilt-by-association method used by the authors of AraNet in associating genes with functions/traits (Lee et al, 2010): the confidence associated with an 'unknown' gene is simply the sum of the weights of the interactions connecting it to genes annotated with the function.

5. Local and Local+: These algorithms are variations of guilt by association that consider the number of edges incident on each unknown example. Local takes negative examples into account whereas Local+ does not.

SinkSource, Hopfield networks, and FunctionalFlow (with more than one phase) attempt to take all the interactions in the network into consideration, while the other approaches only consider direct linkages. SinkSource is similar to the GeneMANIA algorithm for gene function prediction (Mostafavi et al, 2008) with the difference being that, during the prediction process, GeneMANIA allows the confidence scores of the 'known' genes (positive and negative examples) to vary from their original values based on the underlying network.

Typically, the performance of gene function prediction algorithms is measured in terms of the precision (fraction of predictions that are correct) achieved by an algorithm at different values of recall (fraction of positive instances predicted correctly) plotted in the precision-recall curve. Since computational gene function prediction precedes and guides experimental validation, given a ranked list of novel predictions, an experimenter would choose a manageable number of top-scoring predictions to pursue. Such predictions correspond to the high-precision, low-recall regime. Hence, we choose precision at 20% recall (P20R) as the measure of performance and use this throughout the rest of the paper.

Thus, based on the setup depicted in Figure 4.2, for each of the 374 specific functions: i) we partition the positive and negative examples into five sets first, and then form 'seed' and 'test' sets from each $4/5^{th}$-$1/5^{th}$ fraction. This ensures each positive or negative example is used in the test set exactly once; ii) then, using each of the six gene function prediction methods, we make predictions based on a seed set ($4/5^{th}$) and compare the results to its corresponding test set ($1/5^{th}$) to calculate true-positive (TP), false-positive (FP) and false-negative (FN) counts; iii) repeat steps (i) and (ii) for the 5 different seed/test partitions and use the TP, FP and FN counts from the 5 runs to calculate P20R, where P (precision) = (TP/[TP+FP]) and R (recall) = (TP/[TP+FN]). This set of runs constitutes the 5-fold cross-validation scheme. Following this pipeline, for each prediction method, we get 374 P20R values (one per function). We then use this distribution of P20R values (which ranges between 0 and 1) to compare methods to each other. Here, the better

a method is at predicting gene function, the closer to 1 is its median distribution of P20R values. Furthermore, one method can be considered a better predictor than another when the former's median P20R is significantly higher than that of the latter, given their full distributions.

4.3.1. Performance of gene function prediction algorithms for a plant gene network

First, we compare the six gene function prediction methods based on the distribution of P20R scores across the 374 specific functions we have selected. These distributions for all 374 functions or for subsets (see below) are represented as box plots in Figure 4.3 for comparison, with the different methods color-coded differently. Since we have two sets of annotations – 'all ECs' and 'sans ISS IEA' – we have two distributions per method presented next to each other, differentiated based on the thickness of the box. Statistical comparison between algorithms across functions for a given set of annotations (all ECs or sans ISS IEA) is done using the Wilcoxon signed rank test to measure the significance of the difference between means of the distributions. The result of each statistical test is summarized as the $-log_{10}$(P-value), where the lower the Wilcoxon P-value, higher its negative logarithm.

On the first glance of the plot for all 374 functions, it is evident that all algorithms perform rather poorly, with the bulk of each P20R distribution below 0.4 (Fig. 4.3A). However, among them, SinkSource has a significantly better performance than the other algorithms in both EC sets (the first bars in each group in Fig. 4.3E, F), followed by Hopfield as the close second. Considering alongside the observation that Local performs very poorly, these results shows that while it helps to take advantage of negative examples in making predictions based on overall network topology, it hurts to learn from negative examples when using a simple guilt-by-association algorithm (Fig. 4.S1A, Fig. 4.S2A and 4.S2C). In support of this conclusion, the direct neighborhood method that uses only positive examples, Local+, does much better than Local, and almost as well as SinkSource. FunctionalFlow 1-phase is simpler than, but performs nearly as well as Local+. On the other hand, flowing information through the network just using positive examples (FunctionalFlow 6-phases) does not improve upon FunctionalFlow 1-phase and, sometimes, even hurts (see below).

This trend in performance remains the same for both sets of annotations: all annotations (all-ECs) and only annotations related to experimental/expert evidences (sans-ISS-IEA). However, except for SinkSource, the median P20R is always higher for sans-ISS-IEA than all-ECs, indicating that there might be intrinsic factors that affect which type of annotation is useful in making quality predictions. One such potential factor is the size of the function (number of genes annotated to the function). Therefore, instead of just assessing performance across all functions simultaneously, we divide the set of 374 functions into 3 size-based groups of ~125 each for all-ECs and sans-ISS-IEA separately and analyze further. Note that a function with different number of annotated genes based on all-ECs and sans-ISS-IEA can be in one group for all-ECs and in another for sans-ISS-IEA.

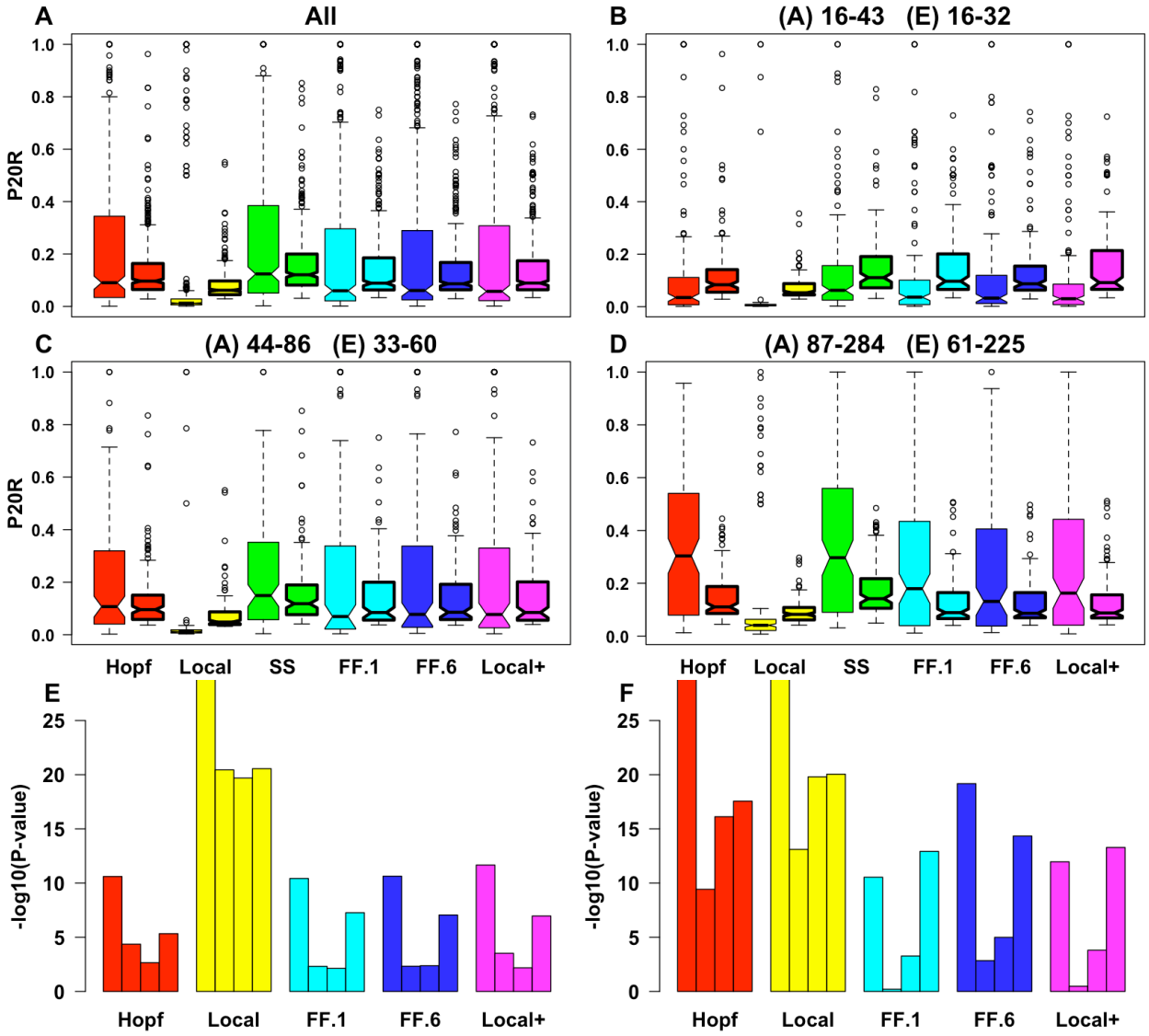


Figure 4.3: Performance of the six gene function prediction algorithms on AraNet measured as precision at 20% recall (P20R) for 374 specific-functions based on all-ECs (A; thin boxes) and sans-ISS-IEA (E; thick boxes) annotations.
The distribution of P20R values for each algorithm is represented as a box plot, where, each box extends from the 25^{th} to the 75^{th} quartile of the distribution with the median (50^{th} quartile) marked in the center of the box. The algorithms are labeled as follows: Hopf: Hopfield, SS: SinkSource, FF.1: FunctionalFlow 1-phase, FF.6: FunctionalFlow 6-phases. **A**) Plots for all 374 functions; **B**, **C** and **D**) Plots for 3 size-based groups containing ~125 functions each. Actual size ranges for 'A' and 'E' in each group are indicated at the top of the plots. Plots **E** and **F** represent the level of significance ($-log_{10}$(P-value)), derived from the P-values of Wilcoxon test comparing the P20R values from SS to that from the other algorithms for all-ECs and sans-ISS-IEA, respectively. In each plot for each algorithm, the first bar corresponds to significance using all 374 functions and the following three bars are significance values calculated within the three size-based groups in increasing order.

For functions with few annotated genes (Fig. 4.3B), performance is poor irrespective of the algorithm. However, learning only from annotations related to experimental evidences (sans-ISS-IEA) leads to better performance than using all annotations (all-ECs), and the algorithms are most equal in the former case without a clear winner (second bars in Fig. 4.3F; first rows in Fig. 4.S1 A and B). For functions with a moderate number of annotated genes (Fig. 4.3C) the results are slightly different. All algorithms pick up in performance (without any one being much better than the others) when using all-ECs (third bars in Fig. 4.3E) and significantly better than using only sans-ISS-IEA. However, using only sans-IEA-ISS, though the performance of all algorithms still remains low, SinkSource gets ahead of others (third bars in Fig. 4.3F; second rows in Fig. 4.S1 A and B). For functions with a large number of annotated genes (Fig. 4.3D) using all annotations is clearly better than using only sans-ISS-IEA and SinkSource performs very significantly well compared to other methods in both cases (fourth bars in Fig. 4.3E and F; first rows in Fig. 4.S1 A and B). Hopfield presents an interesting case here: its performance has been very close to that of SinkSource throughout and yet the difference between these algorithms is statistically significant because of the fact that for most functions, the P20R values from SinkSource have been, although by a little, consistently higher than that from Hopfield (Fig. 4.S1). To provide another perspective of these same results, we plot the number or fraction of functions for which a particular method achieves the maximum P20R (Fig. 4.S2) and draw similar conclusions.

4.3.2. Correlation of performance with network properties

When making function prediction based on an underlying network, it is expected that the topological properties of genes in the network influence the 'predictability' of a function. Among

these are properties calculated based on the network induced by the genes annotated to a function, namely i) the number of genes, ii) the fraction among these genes that form the largest connected component, iii) the number of components, iv) total edge weight, v) weighted density, and vi) average weighted degree (see *Methods* for definitions). Others like 'average segregation' take into account interactions within and across functional cohorts of genes based on the hypothesis that genes interact predominantly with other genes belonging to the same function, and only to a lesser degree with genes belonging to other functions (Chagoyen & Pazos, 2010; Yook et al, 2004). Here we assess the correlation of the six properties of the induced network listed above along with average segregation with predictability of functions by SinkSource.

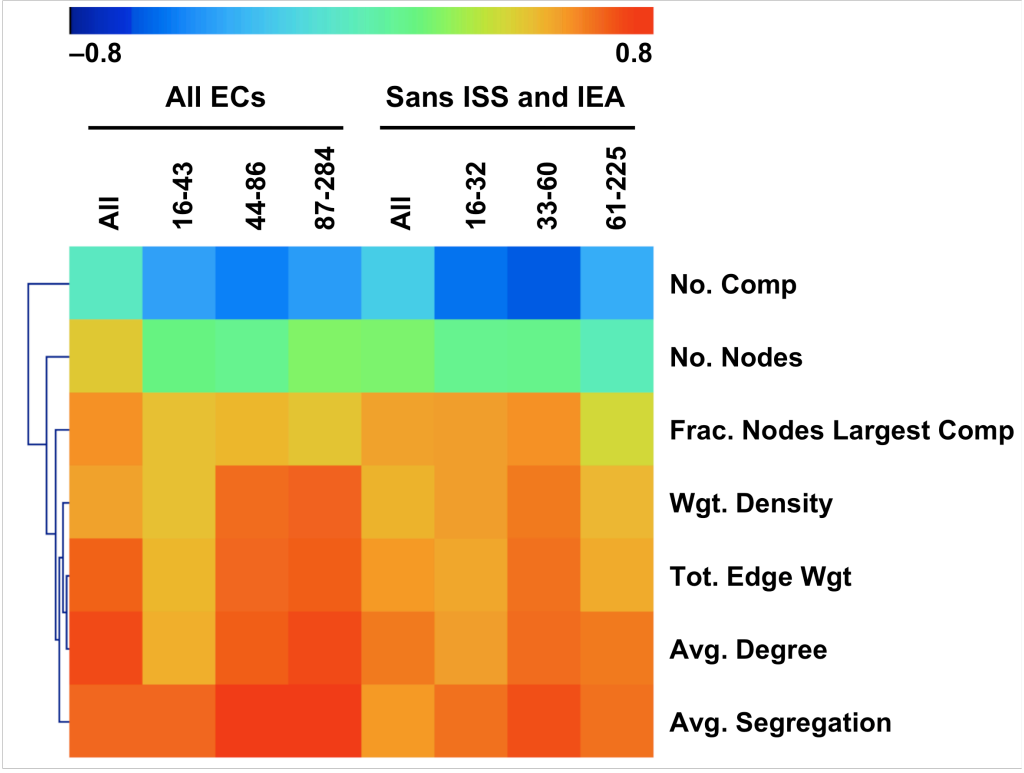


Figure 4.4: Spearman rank correlation of P20R values obtained using the SinkSource algorithm to the topological properties of genes belonging to a function in the original network.
For each function in the context of its induced network, No.Comp: number of connected components; No.Nodes: number of annotated genes in the network; Frac.Nodes.Largest.Comp: fraction of genes in the largest connected component; Wgt.Density: weighted density of edges; Tot.Edge.Wgt: total edge weight; Avg.Degree: average degree of genes; and, Avg.Segregation: average segregation.

We measure the association of a network property to performance using the Spearman rank correlation of the vector of SinkSource P20R values of the functions to the vector of their network property scores (Fig. 4). Interestingly, the number of genes in a function has a reasonable amount of correlation with performance when using all annotations but not when

using only sans-ISS-IEA annotations. Hence, for the rest of the properties, measuring correlation within size-based groups (in addition to all functions) helps to remove any effect of the number of annotated genes (similar to taking a 'partial' correlation). As expected, it is clear (especially within the size-based groups) that the more number of disconnected components the genes are broken down into, the less predictable their function is. A corollary is the positive correlation to the fraction of genes in the largest component. Although, total edge weight and average degree by definition increase as the number of genes in the functions increases (observed as increasing correlation with performance with increasing group size), high correlation within the groups indicates that the connectivity among genes is more important than the number of genes itself. Weighted density displays a similar trend even though it is independent of the number of genes. The strongest correlation to performance, however, comes from the average segregation. Since this measure accounts for the coherence of genes with respect to the connectivity to the rest of the network, it is intuitively the closest to defining functionally related genes in a modular network, making it the favored candidate for determining predictability of functions.

4.3.3. Performance on plant-specific functions

In addition to examining computational methods for predicting gene function, we are also simultaneously evaluating which functions are more and less 'predictable', providing indications about the limits of our knowledge of biological function and/or gene interactions. Specifically, it of interest to ask if there are evolutionary classes of functions that can be predicted with more confidence than others. In the case of network-based function prediction, since the underlying functional interaction network that guides a prediction algorithm is more often than not heavily derived from information in other well-studied species – AraNet being a case in point with data on several types of associations borrowed from human, fly, worm and yeast models – the relevant question is about predicting plant-specific versus conserved functions.

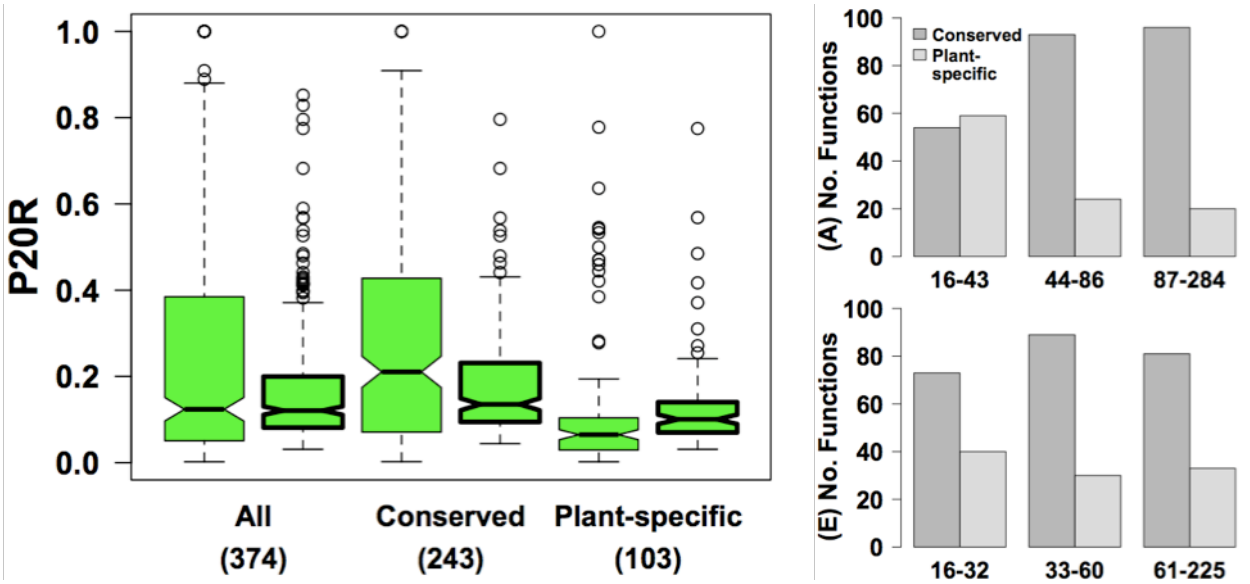


Figure 4.5: Performance (P20R) of SS in 243 conserved and 103 plant-specific functions.
(***Left***) Box plots of P20R values for each group. Results using all annotations are indicated using thin boxes, while those using sans-ISS-IEA annotations are indicated using thick boxes. Adjoining bar plots (***right***) depict the number of plant-specific functions in the 3 size-based groups for all-ECs (A) and sans-ISS-IEA (E).

To address this question, among the 374 functions, we first identify 243 that are conserved and 103 that are plant-specific (leaving behind 28 moderately-conserved functions) based on the number of genes annotated to each function in Arabidopsis compared to the numbers for that function in the human, fly, worm and yeast genomes (see *Methods*; Table 4.S1). Then, on juxtaposing the 'predictability' (P20R) of these classes of functions using SinkSource (Fig. 4.5 *left*), we observe that conserved functions are more predictable than all functions considered together, which is a consequence of the fact that the plant-specific functions are much less predictable. These two classes also differ in the effect of using computational/electronic annotations: while using all annotations is better for conserved functions (Wilcox P-value <1E-7), when considering plant-specific functions, restricting to learn only from experimental/expert annotations aids in greater prediction performance compared to using all annotations (Wilcox P-value =0.000533), although both are poor. Since this relationship is reminiscent of the analysis for different size-based groups of functions (Fig. 4.3), we check the number of genes annotated to conserved- and plant-specific functions. Here, we note that when using all annotations, a majority of the conserved and plant-specific functions fall in the 'medium/large' and 'small' groups, respectively, leading to the poor and good predictability of the functions. However, using

just sans-ISS-IEA, the number of functions is more or less evenly spread among the three groups for both classes, which is hurting the conserved and helping the plant-specific functions.

We use examples among the most and least predictable conserved (Table 4.1) and plant-specific (Table 4.2) functions to gain more understanding. Conserved functions, specifically those concerned with basic processes like protein folding, nucleotide transport, innate immunity, cytoskeleton organization and cell cycle, are generally expected to be highly predictable. However, it is striking that the regulatory functions of several of these basic processes that are still conserved in all species are poorly predictable. On the other hand, cell wall modification, auxin or cytokinin signaling, and photosynthesis are among the most predictable plant-specific functions although within the top few predictions, P20R falls to values about 0.5. The most specialized functions in plants have to do with development, morphogenesis, pattern formation and phase transitions of various tissues, organs or growth stages. It is exactly for these functions that new genes are hardest to predict.

	Function			All ECs				Sans ISS and IEA			
	GO ID	GO Term	Depth	No. Genes	No. Nodes	SS P20R	Avg. Seg	No. Genes	No. Nodes	SS P20R	Avg. Seg
Most Predictable	6457	protein folding	5	228	221	1.000	23.736	25	25	0.177	18.816
	46942	carboxylic acid transport	4	74	73	1.000	51.110	22	21	0.330	1.524
	7031	peroxisome organization	3	25	22	1.000	91.524	25	22	0.796	91.524
	50790	regulation of catalytic activity	3	132	102	1.000	17.493	31	27	0.172	6.847
	65009	regulation of molecular function	2	141	110	0.909	14.003	39	34	0.167	5.344
	15931	nucleobase nucleoside nucleotide and nucleic acid transport	3	39	35	0.889	5.911	38	34	0.567	6.245
	45087	innate immune response	3	255	224	0.880	31.602	136	123	0.143	4.062
	226	microtubule cytoskeleton organization	3	38	36	0.875	25.179	32	30	0.130	14.683
	16137	glycoside metabolic process	4	102	99	0.870	28.655	53	50	0.152	10.074
	55085	transmembrane transport	2	222	217	0.860	12.498	29	26	0.290	26.939
	51726	regulation of cell cycle	4	104	98	0.800	38.154	48	42	0.063	10.126
	19725	cellular homeostasis	2	181	170	0.791	22.490	71	64	0.309	21.328
Least Predictable	42127	regulation of cell proliferation	4	23	18	0.011	1.725	22	18	0.058	1.725
	45088	regulation of innate immune response	5	32	26	0.007	5.493	30	25	0.057	5.730
	45595	regulation of cell differentiation	4	36	31	0.053	6.844	36	31	0.056	6.844
	48645	organ formation	3	32	24	0.006	0.420	32	24	0.056	0.420
	31327	negative regulation of cellular biosynthetic process	6	70	64	0.036	2.902	58	52	0.055	2.890
	45165	cell fate commitment	3	27	21	0.053	13.576	27	21	0.055	13.576
	6997	nucleus organization	3	25	22	0.004	0.000	25	22	0.054	0.000
	45934	negative regulation of nucleobase nucleoside nucleotide and nucleic acid metabolic process	6	65	60	0.047	2.979	55	50	0.053	3.220
	19827	stem cell maintenance	2	24	20	0.042	12.780	24	20	0.052	12.780
	48584	positive regulation of response to stimulus	4	50	42	0.028	3.677	48	41	0.051	3.771
	31324	negative regulation of cellular metabolic process	5	87	81	0.023	2.420	75	69	0.050	2.347
	40008	regulation of growth	3	43	33	0.004	0.751	43	33	0.044	0.751

Table 4.1: Table of the most and least predictable conserved functions using SS in both evidence-code combinations.

	Function			All ECs				Sans ISS and IEA			
	GO ID	GO Term	Depth	No. Genes	No. Nodes	SS P20R	Avg. Seg	No. Genes	No. Nodes	SS P20R	Avg. Seg
Most Predictable	9926	auxin polar transport	4	46	43	1.000	34.554	37	34	0.180	3.281
	9664	plant-type cell wall organization	4	75	72	0.778	41.100	57	55	0.568	2.453
	9736	cytokinin mediated signaling	5	39	36	0.533	212.990	39	36	0.775	212.990
	15979	photosynthesis	3	147	143	0.636	22.377	95	93	0.371	18.620
	42545	cell wall modification	4	129	87	0.545	2.289	67	64	0.485	3.272
	9735	response to cytokinin stimulus	4	75	65	0.542	27.003	75	65	0.417	27.003
	10017	red or far red light signaling pathway	3	28	26	0.500	21.610	28	26	0.191	21.610
	9640	photomorphogenesis	3	46	39	0.471	23.981	46	39	0.169	23.981
	9825	multidimensional cell growth	3	46	44	0.471	5.282	22	20	0.040	9.188
	19684	photosynthesis light reaction	4	84	82	0.459	19.832	71	69	0.310	18.377
	10431	seed maturation	3	22	18	0.444	7.504	21	18	0.170	7.504
	9639	response to red or far red light	5	146	129	0.421	5.588	130	113	0.157	6.272
Least Predictable	10228	vegetative to reproductive phase transition	2	70	61	0.019	3.437	70	61	0.053	3.437
	48653	anther development	3	24	16	0.010	0.000	24	16	0.052	0.000
	9553	embryo sac development	4	76	63	0.008	1.067	76	63	0.049	1.067
	9740	gibberellic acid mediated signaling	5	23	21	0.025	0.000	23	21	0.048	0.000
	10051	xylem and phloem pattern formation	4	40	31	0.014	0.896	39	30	0.047	0.930
	48825	cotyledon development	4	27	18	0.005	0.000	27	18	0.046	0.000
	48573	photoperiodism flowering	3	37	33	0.007	1.814	37	33	0.045	1.814
	9561	megagametogenesis	2	42	33	0.005	2.119	42	33	0.041	2.119
	10102	lateral root morphogenesis	5	25	21	0.002	0.000	25	21	0.036	0.000
	9734	auxin mediated signaling pathway	5	28	21	0.003	1.696	27	20	0.036	1.862
	9567	double fertilization forming a zygote and endosperm	3	24	18	0.004	0.000	24	18	0.034	0.000
	48444	floral organ morphogenesis	4	28	21	0.002	0.000	28	21	0.031	0.000

Table 4.2: Table of the most and least predictable plant-specific functions using SS in both evidence-code combinations.

4.4. Discussion

Gene function prediction is a crucial aspect of post-genomic biology that warrants investment of large-scale efforts to invent and assess novel methods that make the best use of available data and make testable predictions, even in the best studied of model organisms (Peña-Castillo et al, 2008). Network representations of the molecular interactions have become mainstay abstractions that are convenient for modeling cellular organization (Barabasi & Oltvai, 2004). Since a gene or its product engages in interactions with other components to perform a biological function, analyzing the network neighborhood of a gene will provide clues about its functional context.

Arabidopsis is the forerunner as a model for plants with a genome sequenced a decade ago and genomic data amassing at an ever-increasing rate. Nonetheless, associating genes with specific functions and high-level traits/phenotypes is a surviving challenge with half the genes in its genome not annotated with any function. When several attempts are being made at figuring out a genome-scale map of Arabidopsis gene interactions (Alexeyenko & Sonnhammer, 2009; Cui et al, 2008; Ma et al, 2007), now is a prime time for assessing how useful such maps are in aiding gene function prediction.

We evaluate the performance of a gamut of algorithms in predicting gene functions in Arabidopsis based on an underlying network of gene interactions, AraNet (Lee et al, 2010). AraNet is a recent effort that represents the conglomeration of network-level genomic data from the major model organisms – yeast, fly, worm and human – put to good use in predicting gene associations in a plant model, meagerly supplemented with data from the plant itself. Due to its large-scale integrative nature and public availability, we consider this network as a current draft of the Arabidopsis interactome.

Lee et al. (2010) evaluate the performance of AraNet in inferring gene function using a local algorithm (*here*, FunctionalFlow 1-phase). They measured prediction performance after leave-one-out cross-validation using the area under the Receiver Operating Characteristic (ROC) curve. Our analysis differs from theirs in the following two aspects:

a) We use *k*-fold cross validation with a small *k* (*k*=5) to better reflect the reality that only a fraction of the genes annotated to a function are already known. In contrast, leave-one-out cross validation erases one positive example at a time, a procedure that may lead to the over-estimation of prediction performance.
b) We calculate precision for different values of recall instead of tracing the relationship between the false positive rate and true positive rate (recall) as in a ROC curve. The rationale is that the number of negative examples is typically much larger than the number of positive examples for any given function. Therefore, a large change in the number of false positives may cause only a small change in the false positive rate used in ROC analysis. In such situations, precision-recall curves are better indicators of performance than ROC curves (Jansen & Gerstein, 2004). Precision compares the number of false positives to the number of

predictions, as opposed to the false positive rate, which compares the number of false positives to the much larger number of negative examples. Therefore, a precision-based measure rewards high-quality positive predictions without regard to the accuracy of negative predictions, which are far less helpful in guiding wet-lab experiments.

A main concern with the evaluation method used by Lee et al. (2010) is that the evaluation on prediction of GO BP terms (functions) is performed using the final network that was learnt from a gold-standard based on the same GO BP annotations: a procedure that is circular and could lead to over-estimation of performance. Because we only have access to the final network, we would like to draw attention to the fact that we too use this integrated network in performing the evaluations and thus face the same over-estimation problem. However, since we concentrate on discovering trends and reasoning through distributions of scores (not exact calculations and novel predictions), we maintain that our results hold.

Using a semi-supervised learning and evaluation scheme (Figure 2) involving 5-fold cross-validation, we find that algorithms (Hopfield, SinkSource) that use large-scale network topology perform better than others that rely only on local neighborhood. Overall, SinkSource makes the best predictions of Arabidopsis gene function using AraNet. A lot more is revealed when we examine performance of the algorithms on functions annotating small, moderate or large numbers of genes. All algorithms perform poorly when only a small number of genes are 'known' for a particular function or when the function is very specific. When we further dissect the annotations of genes to functions based on the source of annotation, we see that using only annotations based on experimental/expert evidences is better than using all annotations in making confident predictions for the small-size functions. We therefore suggest, when seeking candidate genes with poorly/sparsely-annotated functions, to use only experimentally verified annotations to make new predictions. On the other hand, when a considerable number of genes are annotated to a function, we suggest using all annotations including computational and electronic evidences to make new predictions.

Since the underlying network is the basis of association between the genes, several network properties of genes annotated with a function influence the predictability of that function. The

property most correlated with function prediction performance is average segregation that measures how well connected are genes to other genes in the same function and how separated they are with respect to rest of the network. This property is akin to the notion of functional modularity of biological networks (Hartwell et al, 1999). Given the property is, by definition, independent of the number of genes annotated to a function, there is a clear trend where functions with greater and greater number of annotated genes also have higher and higher average segregation from the rest of the genes. Since this property correlates well with performance, we extend that small-size functions are poorly segregated in AraNet leading to poor prediction.

As mentioned earlier, since data from other model organisms dominates AraNet, it is vital to test how well the network performs in predicting plant-specific versus conserved functions. We observe that plant-specific functions are hardly predicted well while conserved functions are mostly much more predictable. Given that many poorly predicted functions have small number of annotated genes, it is intriguing that several conserved functions that contain reasonably large number of genes are still not predicted well. Many of these functions being related to regulation of biological processes hints to the fact that while different organisms contain a machinery to control a conserved function, how this regulation is achieved in terms of the exact genes involved might be very different causing the same regulatory function in different species to contain non-homologous genes. In other words, regulation of several functions have become 'specialized' in different species in order for the same target function (e.g. innate immunity) to be regulated by very different stimuli that different organisms encounter.

The results presented above indicate the following: i) conserved functions can be predicted very well compared to plant-specific functions, which are rarely well-predicted; ii) this points to a large gap in our knowledge of plant-specific gene interactions that is hard to fill when banking heavily on data from other species while reconstructing a genome-scale network; ii) computational methods for gene function prediction that borrow information from genes of 'similar' function in deciding on a given function for a new gene should be developed, taking forward some efforts already in place (Mostafavi & Morris, 2010; Pandey et al, 2009); and iii) as we can only reap as much as we sow, it is important to perform more experiments and explore

the gene space of poorly studied functions to work towards informative experiments that can be brought into the network generation and prediction pipeline. Several such plant-specific functions and 'specialized' conserved functions are provided in this study (Table 4.S1) that point to avenues for future research in plants.

4.5. Methods

We use Gene Ontology (GO) biological process (BP) term annotations to define gene function. GO functional annotation for Arabidopsis was downloaded from TAIR (Swarbreck et al, 2008) in October 2009 and the GO annotation hierarchy was downloaded from the GO database in July 2009. We extracted the annotations from the 'biological process' hierarchy and applied the true-path-rule based on 'is_a' and 'part-of' relationships between GO terms, whereby genes annotated with a given term are also annotated with all its ancestors. From all the terms in the BP hierarchy, we first define 'specific' terms as those that annotate <300 genes (~1% of the genes in the genome). We filter the terms further by choosing only those that annotate 20 or more genes to ensure that there are enough genes for cross-validation (see below). Finally, we filter out parent terms in every pair of parent-child terms in the GO DAG that differ in the number of annotated genes by <=3 in order to avoid obviously highly overlapping GO terms, leaving a sets 374 specific GO BP terms (functions) to work with (Table 4.S1). We create two sets of gene-function annotations: one including annotations based on all evidence, and another based only on experimental/expert evidences obtained by excluding annotations solely based on sequence/structural similarity (ISS) or electronic annotation (IEA).

We downloaded AraNet v1 in February 2010. Using AraNet, we applied the semi-supervised learning scheme represented in Figure 2 for each of the 374 functions and evaluated the performance of six network-based gene function prediction algorithms (described below) using 5-fold cross-validation. Software implementing the function prediction algorithms is available at http://bioinformatics.cs.vt.edu/~murali/software/gain.

4.5.1. The SinkSource Algorithm

We model the gene functional interaction network as an undirected graph $G=(V, E)$, consisting of a set V of nodes (i.e., genes) and a set E of edges (i.e., functional interactions). Let w_{uv} denote the weight of the edge $(u, v) \in E$, denoting the log-likelihood score of the interaction defined in (Lee et al, 2010). For each function f, we partition V into three subsets, V^+, V^0 and V^- as follows: V^+ is the set of nodes annotated with f (after applying the true-path-rule; positive examples), V^- is the set of nodes not annotated to f or to any of its ancestors (negative examples), and V^0 is the remaining set of nodes. For each node $v \in V^0$, our goal was to assess whether v should be a member of V^+ or V^-. We do so by computing a function $r : V \rightarrow [0,1]$ that is smooth over G. Specifically, we set $r(v) = 1$ for every node $v \in V^+$, $r(v) = 0$ for every node $v \in V^-$, and required that r minimize the function

$$S(G,r) = \sum_{(u,v)\in E} w_{uv}\left(r(u) - r(v)\right)^2$$

Minimizing $S(G, r)$ enforces the smoothness of r in the sense that the larger the weight of an edge (u, v), the closer in value $r(u)$ and $r(v)$ must be. The function $S(G, r)$ is minimized when

$$r(v) = \frac{\sum_{u\in N_v} w_{uv} r(u)}{\sum_{u\in N_v} w_{uv}}, \qquad (1)$$

where N_v is the set of neighbors of node v (Zhu et al, 2003). The right-hand side of this equation can be split into two parts: one corresponding to contributions to $r(v)$ from neighbors in V^0 and the second to a constant contribution from neighbors in V^+ and V^-. Let r^0 denote the vector of values taken by the function r at the nodes in V^0. Let M denote the square matrix, where $M_{uv} = w_{uv} \Big/ \sum_{v\in N_u} w_{uv}$, for every $u, v \in V^0$. We see that r^0 satisfies the equations $r^0 = M r^0 + c$, where c is a vector denoting contributions from V^+ and V^-. We compute r^0 by initializing it to 0 for each node $u, v \in V^0$ and repeatedly applying the operation $r^0 = M r^0 + c$. This process is known to converge (Zhu et al, 2003), yielding a value of $r^0 = (I - M)^{-1} c$, where I is the identity matrix. The matrix M is sparse, being the adjacency matrix of a functional interaction network. Therefore, this iterative approach is efficient in practice.

4.5.2. Other Algorithms

We implemented five other algorithms for the purpose of comparison. Of these, two algorithms use both positive and negative examples. The other three algorithms do not use negative examples for making predictions. We use both types of algorithms in order to assess the impact of negative examples on the cross-validation results.

a) The Hopfield network algorithm (Karaoz et al, 2004) sets $r(v) = 1$ for every node $v \in V^+$, $r(v) = -1$ for every node $v \in V^-$, and initializes $r(v) = 0$ for every node in $v \in V^0$. The algorithm repeatedly applies a modified form of equation (1), by setting $r(v)$ to be the sign of the right hand side of equation (1). Thus, it restricts $r(v)$ to take the value 1 or -1. This process is also known to converge.
b) The FunctionalFlow algorithm (Nabieva et al, 2005) does not use negative examples. Each positive example has an infinite reservoir of fluid. The algorithm runs in phases. In each phase, fluid flows along each edge from the node with a larger reservoir to the node with a smaller reservoir. Please see the original paper for the precise equations governing the flow. The total inflow into a node over all phases represents the confidence with which the node is predicted with a function. This algorithm needs the number of phases as input. As suggested by the authors, we used half the diameter of the AraNet network (the diameter was 12, so we used 6 phases). In addition, we run this algorithm for one phase, in which case it is equivalent to the guilt-by-association method used by Lee et al (2010).
c) The Local algorithm (also called guilt-by-association in the literature) initializes $r(v) = 0$ for each node $v \in V^0$ and applies equation (1) exactly once to each node $v \in V^0$. While this form uses both positive and negative examples, a variation of this algorithm, Local+, uses only positive examples.

Although Local+ and FunctionalFlow (1 and 6 phases) do not use negative examples when making predictions, we use negative examples when computing the performance of these algorithms on cross-validation.

We use a set of topological properties to characterize the genes annotated with a function in the network: In the context of the whole network $G=(V, E)$, consisting of a set V of nodes (i.e.,

genes) and a set E of edges (i.e., functional interactions), we first define the graph $G_f=(V_f, E_f)$ as the graph induced by the genes annotated with function f, and compute the following measures:

a) Number of nodes $|Vf|$.

b) Number of connected components in G_f where a connected component is defined as a subgraph in which any two vertices are connected to each other by paths, and to which no more vertices or edges (from Gf) can be added while preserving its connectivity.

c) Fraction of nodes in the largest connected component defined as $|V_{f,C}|/|V_f|$ where $V_{f,C}$ is the set of nodes in the largest connected component.

d) Total edge weight, defined as:

$$\sum_{(u,v)\in E_f} w_{uv}$$

e) Average weighted degree, defined as:

$$\frac{\sum_{v\in V_f}\sum_{u\in V_f} w_{uv}}{\left|V_f\right|}$$

f) Weighted density, defined as:

$$\frac{2*\sum_{(u,v)\in E_f} w_{uv}}{\left|V_f\right|\left(\left|V_f\right|-1\right)}$$

g) Average segregation, defined as:

$$\frac{\sum_{(u,v)\in E_f} w_{uv} \Big/ \left|E_f\right|}{\sum_{u\in V_f; y\in V; (u,y)\in E} w_{uy} \Big/ \left|E_{f'}\right|}$$

where $E_{f'}$ is the set of edges in G that are incident on V_f.

For analysis of conserved and plant-specific functions, GO annotations for humans (Hs), fly (Dm), worm (Ce) and yeast (Sc) were downloaded from the GO database in June 2010, and complete annotations for gene annotations for the 374 functions under consideration were created for each species by applying the true-path-rule. First, for each of the 374 functions, we calculate the fraction of the genes in the genome (in each species) that is annotated to that

function. Then, we contrast the fraction of genes annotated with a function (frac) in Arabidopsis (At) to that in other species using an odds-ratio-like measure: (fracAt)/(fracHs + fracDm + fracCe + fracSc + 0.00001). The small number in the denominator helps avoid division by zero when no genes are annotated to a function in all the other species. Plant specific functions are defined as those with a ratio >10 and conserved functions are defined as those with a ratio <1 to obtain a clear partitioning of functions. This choice leaves 28 functions among the 374 that are 'moderately conserved' and hence not used in determining properties of either group of conserved or plant-specific functions.

All data processing was done using Perl scripts. Statistical analysis and plotting were carried out using R (Ihaka & Gentleman, 1996).

4.6. References

Alexeyenko A, Sonnhammer EL (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res* **19:** 1107-1116

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29

Barabasi A-L, Oltvai Z (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5:** 101-113

Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y (1998) Predicting function: from genes to genomes and back. *J Mol Biol* **283:** 707-725

Chagoyen M, Pazos F (2010) Quantifying the biological significance of gene ontology biological processes--implications for the analysis of systems-wide data. *Bioinformatics (Oxford, England)* **26:** 378-384

Costello J, Dalkilic M, Beason S, Gehlhausen J, Patwardhan R, Middha S, Eads B, Andrews J (2009) Gene networks in Drosophila melanogaster: integrating experimental data to predict gene function. *Genome biology* **10:** R97

Cui J, Li P, Li G, Xu F, Zhao C, Li Y, Yang Z, Wang G, Yu Q, Shi T (2008) AtPID: Arabidopsis thaliana protein interactome database--an integrative platform for plant systems biology. *Nucleic Acids Res* **36:** D999-1008

Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, Troyanskaya OG (2008) A genomewide functional network for the laboratory mouse. *PLoS computational biology* **4:** e1000165

Hartwell L, Hopfield J, Leibler S, Murray A (1999) From molecular to modular cell biology. *Nature* **402:** C47-C52

Hu P, Janga SC, Babu M, Diaz-Mejia JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A et al (2009) Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol* **7:** e96

Huttenhower C, Haley E, Hibbs M, Dumeaux V, Barrett D, Coller H, Troyanskaya O (2009) Exploring the human genome with functional maps. *Genome research* **19:** 1093-1106

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. *J Comput Graph Stat* **5:** 299-314

Jansen R, Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol* **7:** 535-545

Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci U S A* **101:** 2888-2893

Koornneef M, Meinke D (2010) The development of Arabidopsis as a model plant. *Plant J* **61:** 909-921

Lee I, Ambaru B, Thakkar P, Marcotte E, Rhee S (2010) Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. *Nature biotechnology* **28:** 149-156

Lee I, Lehner B, Crombie C, Wong W, Fraser A, Marcotte E (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nature Genetics* **40:** 181-188

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

Mostafavi S, Morris Q (2010) Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* **26:** 1759-1765

Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology* **9 Suppl 1:** S4

Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21 Suppl 1:** i302-310

Pandey G, Myers C, Kumar V (2009) Incorporating functional inter-relationships into protein function prediction algorithms. *BMC Bioinformatics* **10:** 142

Peña-Castillo L, Tasan M, Myers C, Lee H, Joshi T, Zhang C, Guan Y, Leone M, Pagnani A, Kim WK, Krumpelman C, Tian W, Obozinski G, Qi Y, Mostafavi S, Lin GN, Berriz G, Gibbons

F, Lanckriet G, Qiu J et al (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome biology* **9 Suppl 1:** S2

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. *Molecular systems biology* **3**

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

Yamada T, Bork P (2009) Evolution of biomolecular networks ‚Äî lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology* **10:** 791-803

Yook SH, Oltvai ZN, Barabasi AL (2004) Functional and topological characterization of protein interaction networks. *Proteomics* **4:** 928-942

Zhu J, Zhang B, Smith E, Drees B, Brem R, Kruglyak L, Bumgarner R, Schadt E (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics* **40:** 854-861

Zhu X, Ghahramani Z, Lafferty J (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *ICML-03, 20th International Conference on Machine Learning*.

5. A transcriptional regulatory network coordinating activation of cellulose and repression of lignin biosynthesis pathways in rice

Arjun Krishnan, Madana Ambavaram, Kurniawan Trijatmiko, Andy Pereira

Submitted for review, 2010.

5.1. Abstract

Cellulose from plant biomass is the largest renewable energy resource of carbon fixed from the atmosphere, which can be converted into fermentable sugars for production into ethanol. However, the cellulose present as lignocellulosic biomass is embedded in a hemicellulose and lignin matrix from which it needs to be extracted for efficient processing. Expression of an *Arabidopsis thaliana* transcription factor SHINE (SHN) in rice (*Oryza sativa*), a model for the grasses, causes an increase in cellulose and a reduction in lignin content, predicted based on gene expression and confirmed experimentally. Supporting experiments show that rice SHN lines also exhibit an altered lignin composition correlated with improved digestibility, with no compromise in plant strength and performance. Using a detailed systems-level analysis of global gene expression in rice, we reveal the SHN regulatory network coordinating down-regulation of lignin biosynthesis and up-regulation of cellulose and other cell wall biosynthesis pathway genes. The results thus support the development of non-food crops and crop wastes with increased cellulose, and low lignin with good agronomic performance that could improve the economic viability of lignocellulosic crop utilization for biofuels.

5.2. Introduction

Crop residues are a vast resource of lignocellulose feedstock available for conversion to biofuels, and their utilization does not compete with food supplies unlike grain-based feedstocks (Haigler et al, 2001). Rice straw itself constitutes half the crop-waste worldwide, which is either burnt or wasted (Sticklen, 2006). Non-food perennial grasses such as switchgrass and Miscanthus as well as fast growing woody crops make up the bulk of lignocellulosic resources. In either case, plant lignocellulosic cell walls are quite resistant to digestion of the complex polysaccharides

(cellulose) into simple sugars before fermentation due to the presence of heavily cross-linked lignin. Methods to lower lignin and improve the availability and levels of cellulose are therefore important to make the conversion into biofuels economically feasible.

Cellulose is the most abundant biopolymer on earth, comprising 25-50% of plant biomass with an estimated 100 billion tons synthesized annually as a result of photosynthesis (Haigler et al, 2001; Sticklen, 2006). Cellulose is made up of glucose units and is synthesized at the plasma membrane by the cellulose synthase complex, comprising multiple CESA proteins that belong to multigene families in plants (Somerville, 2006). Long chain cellulose polymers are organized into microfibrils that make up the core content of plant cell walls, contributing to the strength, structure and development of plants (Sticklen, 2006). Hemicelluloses are polysaccharides in plant cell walls, and are synthesized by glycosyltransferases (GTs) located in the Golgi membranes. The most important biological role of hemicelluloses is their contribution to strengthening the cell wall by interaction with cellulose and, in some cell walls, with lignin (Scheller & Ulvskov, 2010). Despite its importance, the details regarding the synthesis of hemicelluloses remain very elusive and very little is known about the regulation of the cellulose biosynthesis pathway.

Lignin, the second most abundant polymer, is a complex comprised of guaiacyl (G), syringyl (S) and p-hydroxylphenyl (H) phenylpropanoid units (Fig. 5.S1), contributing to lignin heterogeneity (Boerjan et al, 2003). Angiosperm dicot lignin is primarily composed of G and S units, and monocot lignin is a mixture of G, S and H units (Fig. 5.S1). Among these, the G lignins (found characteristically in abundance in softwoods of gymnosperms like pines) are more resistant to chemical degradation, making the composition of lignin (the relative ratio of G to S units), along with its quantity, crucial for the digestibility of crops for conversion into biofuels and cellulosic products. The monolignol biosynthetic genes – especially PAL, 4CL and CAD genes – have therefore been used in engineering lignin content and composition in several plants (Vanholme et al, 2008). Many of these studies were first reported in non-feedstock model dicot plants such as tobacco and Arabidopsis (Zhou et al, 2009), and the expectation is that similar approaches can be applied to cellulosic feedstock crops, but very few detailed engineering studies have been reported in the grasses, which are a major lignocellulosic resource.

In the grasses, the maize and sorghum brown midrib mutations (Li et al, 2008) show alterations in lignin content and digestibility; the maize *bk2* and rice *bc1* mutations of a similar gene have a brittle phenotype due to reduction in cellulose and cell wall composition with no compensatory changes in lignin (Li et al, 2003); and, the rice *flexible culm 1 (fc1)* mutant has reduced lignin, H and G residues (Li et al, 2009). However, significant reductions in lignin or digestibility in the monocot crops, including the *brown midrib* and other mutants, are also accompanied by reductions in plant growth, biomass, stalk strength, or pathogen resistance (Li et al, 2008).

Several transcription factors (TFs) have also been shown to affect cellulose and lignin content and composition (Kubo et al, 2005; Mele et al, 2003; Zhong et al, 2006; Zhong & Ye, 2009). Transcriptional regulation is achieved by top-level NAC TFs (SND1, NST1/2, VND6/7) that activate a nexus of intermediate TFs, mostly MYBs. These intermediate TFs in turn activate low-level MYB TFs that bind to and activate target cell wall biosynthetic genes, presumably, achieving high levels of specificity. Such a multi-layered regulatory network that affects multiple target genes offers the cell a robust mechanism to achieve coherent changes in the flux through the pathways. Additionally, the network also points to the possibility of existence of multiple knobs and switches that can be tuned to execute specific regulation of different cell wall pathways in order to optimize secondary cell wall composition. Specifically, cell wall with increased cellulose content in combination with reduced lignin for enhanced sugar and ethanol would yield valuable cellulosic feedstock (Jakob et al, 2009).

The Arabidopsis *SHINE (SHN/WIN)* clade of 3 genes, belonging to the AP2/ERF TF family, was previously shown to be involved in wax/cutin lipid regulation and drought tolerance in Arabidopsis (Aharoni et al, 2004; Broun et al, 2004; Kannangara et al, 2007). A homolog *Nud* has also been found in barley, which is responsible for adhesion of the hulls covering the barley seed (probably mediated by a lipid layer), an important trait in the domestication process (Taketa et al, 2008). Following our findings in Arabidopsis, the *SHN2* gene (Aharoni et al, 2004) under control of the CaMV35S promoter was transformed into rice and shown to confer drought resistance and enhanced water use efficiency with a slight increase in cuticular wax (Karaba, 2007). In the present study, we describe our discovery of a novel function of the SHN gene as a

master regulator of lignin and cellulose/cell wall biosynthesis pathways, coordinating down-regulation of lignin biosynthesis and up-regulation of cellulose and other cell wall biosynthesis pathway genes.

5.3. Results

5.3.1. Expression of Arabidopsis *SHN* Gene in Rice Causes Coordinate Regulation of Cell Wall Biosynthetic Genes

Gene expression analysis of the rice SHN lines using Affymetrix GeneChips revealed coordinate regulation of cell wall biosynthesis genes (as annotated in rice by Yokoyama and Nishitani (2004)), with a distinctive up-regulation of cellulose and other cell wall biosynthesis genes, and down-regulation of lignin biosynthesis genes, including PAL, 4CL, HCT, CCR and CAD, along with regulation of key NAC and MYB TFs (Fig. 5.1 and Table 5.S1). Since we observed that the expression of homologous members of large gene families were being altered in the SHN lines, we sought to reliably quantify the expression levels of individual genes in these multi-gene families (e.g. 4CL, CAD and CSEA). The rice Affymetrix GeneChip probesets were therefore reannotated (described in *Methods*) to distinguish members of gene families to the extent that was possible with the available probes. This reannotation can distinguish 35,161 rice gene-based probesets and the corresponding genes, and thus characterize the differential expression of many cell wall pathway genes.

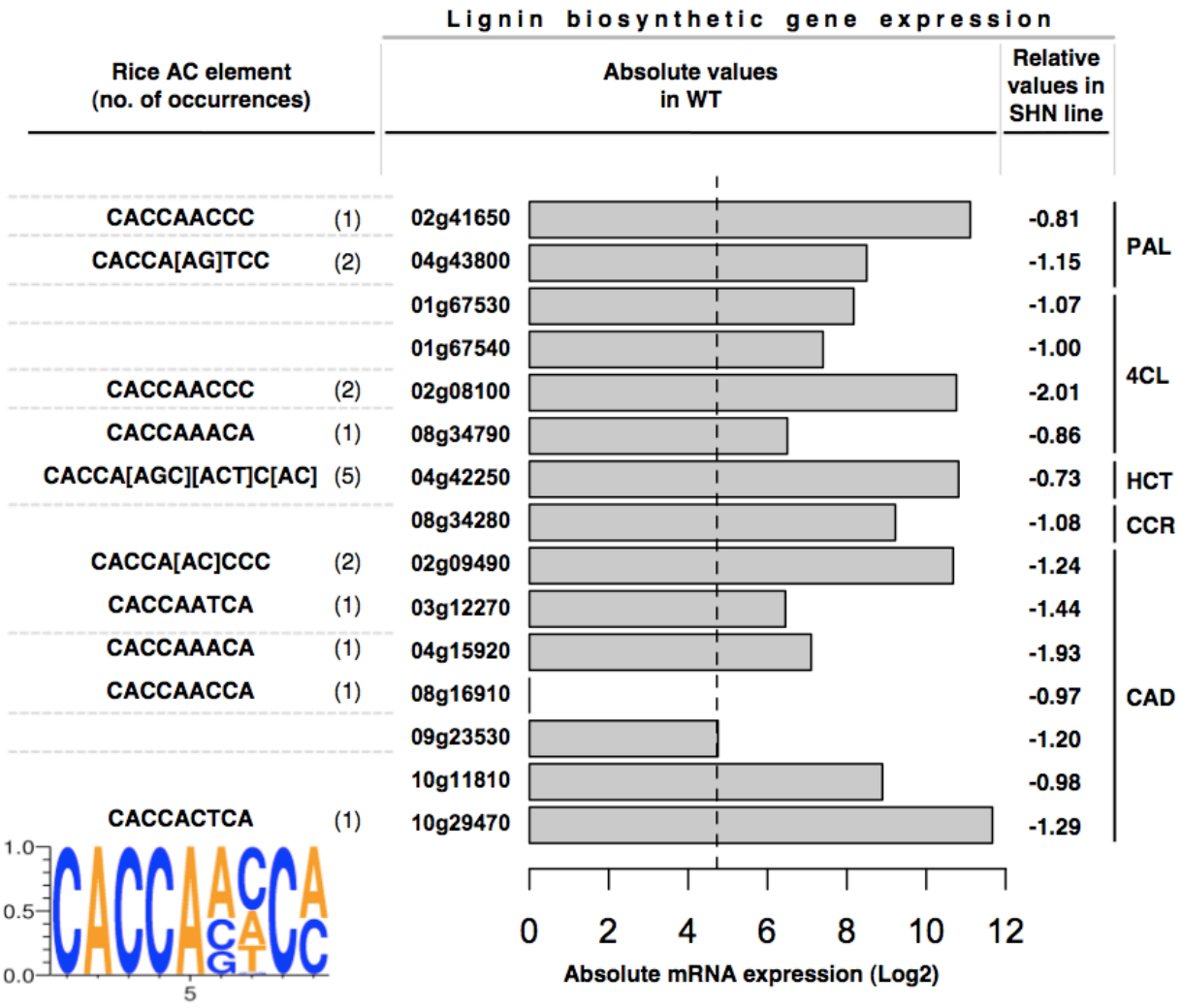


Figure 5.1: Gene expression and regulation of lignin biosynthetic pathway genes.
WT absolute expression levels of SHN-regulated lignin biosynthetic genes obtained from the microarray are plotted in log_2 scale with the 40th percentile of genome-wide gene expression values (marking the level above which genes can be considered to be 'expressed') indicated by a dashed line. The differential expression values of specified genes in SHN lines are shown alongside. The rice AC element sequences identified in the promoters (1 Kb upstream of start site) of the SHN-regulated lignin pathway genes are shown with the number of occurrences of the element in each promoter given in brackets. The sequence logo of the rice AC element identified in the promoter regions of rice lignin biosynthetic genes is given below, with the height of a character at a particular position representing the fraction of occurrences of the corresponding nucleotide in that position.

Repression of the lignin pathway was exhibited as a moderate reduction in expression levels of many enzymes spread across the pathway rather than a drastic reduction of a few enzymes (Fig. 5.1). The moderate reduction in terms of the absolute expression levels indicate that gene expression is not completely shut off, but allows a background flux through the pathway. Moreover, this pathway-wide repression including 4CL and CAD genes that catalyze specific branches of lignin biosynthesis leading to different lignin monomers suggests possible alterations

in lignin composition along with overall reduction in response to SHN expression. Several TFs observed to be regulated by SHN in the microarray experiment (see Table 5.S1) included rice homologs of several transcriptional activators of the cell wall biosynthetic pathways uncovered in *Arabidopsis* and other plant species (Zhong & Ye, 2009). These homologous TFs were hence hypothesized to be the regulators of secondary cell wall biosynthesis in rice, and mediators of the coordinate regulation of the biosynthetic pathways by SHN.

To glean support for similarity of transcriptional regulation of the cell wall pathway genes in rice to that in other species, we performed de novo sequence motif discovery on the promoter regions of the lignin and other cell wall genes. The 1 Kb upstream sequences of the cell wall biosynthetic genes and that of the rest of the genes in the genome were divided into two classes and small DNA motifs in this region that had significantly high association with the former class were recovered. This analysis led to the identification of a rice AC element 'CACCA[ACG]NC[AC]' (Fig. 5.1) that is similar to the ACII element 'CACCAACCC' known to mediate vascular tissue-specific expression of the cell wall biosynthetic genes in other plant species, with evidence for MYB TFs binding to this element (Zhong & Ye, 2009). This suggested that a similar transcriptional machinery of MYB TFs (downstream of NACs) exists in rice, and that SHN is capable of regulating genes involved in lignin and cell wall biosynthesis, possibly by regulating these TFs.

To verify SHN-regulation of the monolignol and cell wall biosynthesis pathways as observed in the microarray, we carried out rigorous quantitative real-time PCR (qRT-PCR) experiments to determine the abundance and tissue-specificity of biosynthetic genes at two different developmental stages – leaf and culm – in SHN and WT plants. We found that transcript levels of eight CAD's, and four 4CL's were significantly repressed in two developmental stages/tissues (Fig. 5.2 *top*), confirming the regulation of lignification in leaf and stem by SHN. Five out of eleven *CesA* genes in rice are significantly up-regulated by SHN in culm tissue (Fig. 5.2 *bottom*). Induction of three out of the five *CesA* genes is consistent across the two developmental stages, leaf and culm, suggesting a role of SHN in inducing the expression of certain cellulose and other cell wall biosynthetic genes in rice. It is worthwhile to note here that although the microarray aided in narrowing our focus on the cell wall pathways, few of the CADs and the CESA genes

confirmed to be differentially expressed using qRT-PCR were not found to be significantly perturbed in the microarray. This dichotomy between the microarray and the experimental results yet again emphasizes the importance of collaborative back-and-forth between experimental and computational analysis.

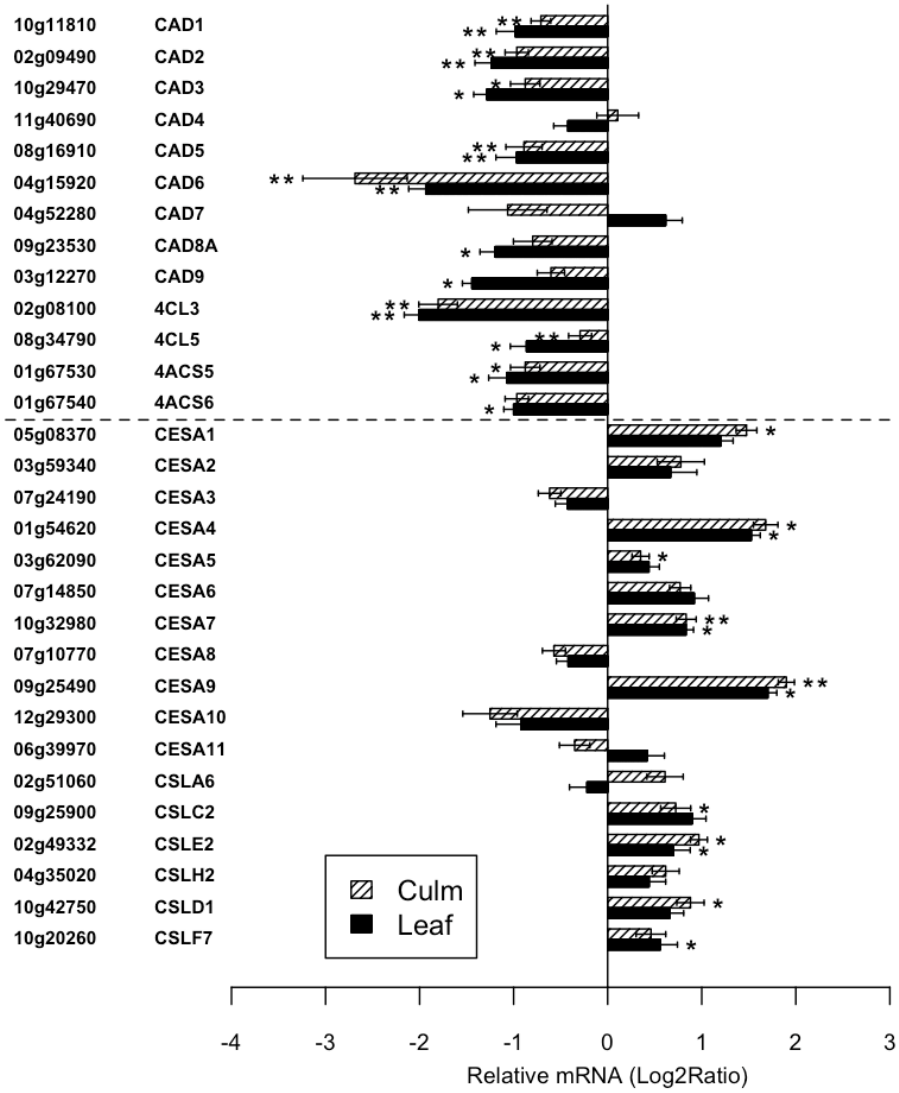


Figure 5.2: qRT-PCR expression analysis of lignin and cellulose biosynthetic genes in SHN leaf and culm.
Data are expressed as the mean relative transcript levels in SHN lines compared to that of WT (log_2 ratio) at each stage (leaf and culm). Error bars represent ± s.e.m. (n=3) (three WT and three SHN lines). Asterisks indicate significant differential expression (t-test; *, P ≤0.05; **, P ≤0.01).

In addition, seven putative rice secondary wall biosynthesis TFs – three NAC genes and five MYB genes – were confirmed to be down-regulated in the leaf tissue (Fig. 5.3). Six of these TFs were also repressed in culm. One TF, MYB20/43 (Os02g49986), was however up-regulated in

both developmental stages (Fig. 5.3). These modes of differential gene expression of TFs may account for the coordinate expression of their biosynthetic targets in new spatial or temporal patterns as required to generate functional integrity of the pathway.

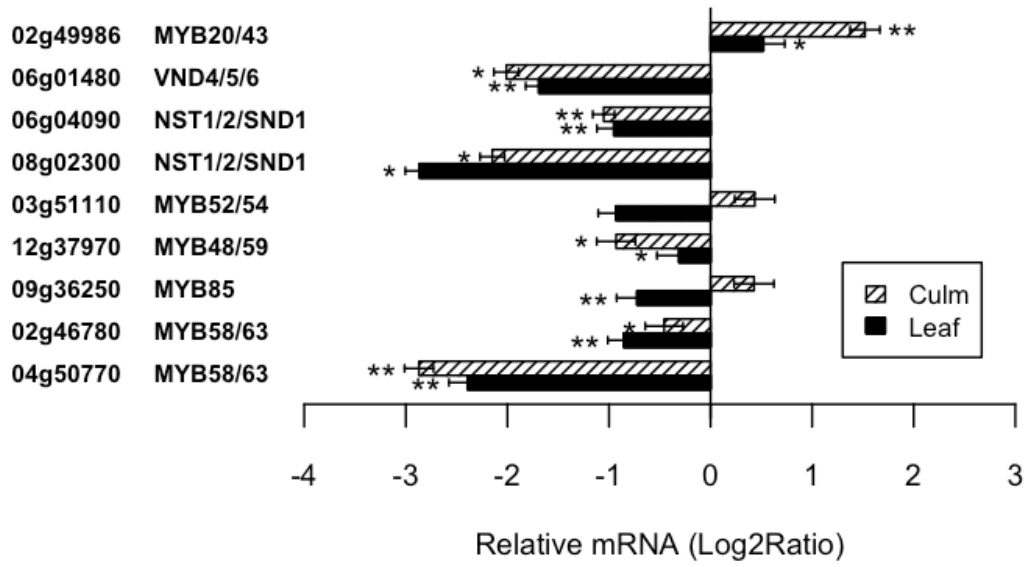


Figure 5.3: Expression analysis of putative secondary cell wall TF genes in SHN leaf and culm.
Data are expressed as the mean relative transcript levels in SHN lines compared to that of WT (*log2* ratio) at each stage (leaf and culm). Error bars represent ± s.e.m. (n=3) (three WT and three SHN lines). Asterisks indicate significant differential expression (t-test; *, P ≤0.05; **, P ≤0.01).

These gene expression results urged us to test for associated phenotypic and biochemical changes. Using extensive confocal microscopy and biochemical analyses, rice SHN lines have been found to have thickened cell walls and uncompromised strength compared to WT (data not shown). Moreover, these SHN lines also have enhanced cellulose and reduced lignin (data not shown). Taken altogether, these data indicated that AtSHN is a key regulator of monolignol and other cell wall biosynthesis.

5.3.2. Transcriptional Network Regulating Lignin and Cellulose Biosynthesis

In order to see if the rice *SHN* gene Os06g40150 (*OsSHN*), homolog of *Arabidopsis SHN2*, also has an intrinsic association with the cell wall pathways, an extensive analysis of coexpression in rice was undertaken. A global coexpression network of rice genes was constructed based on public gene expression datasets in rice. Raw Affymetrix rice expression profiles pertaining 'response to environmental conditions' were collected from GEO (Barrett et al, 2009) and

ArrayExpress (Parkinson et al, 2009). Gene expression values across the diverse conditions were extracted using the custom gene-centric probeset definition, and pairwise gene-gene correlations were calculated to create the global network (see *Methods*). From this global network, the connectivity between i) *AtSHN*-regulated lignin and other cell wall-related genes, ii) TFs associated with these pathways, and other TF genes differentially expressed in the AtSHN microarray, and iii) the *OsSHN* gene, was mined to establish a rough rice transcriptional network of cell wall biosynthesis (Fig. 5.4).

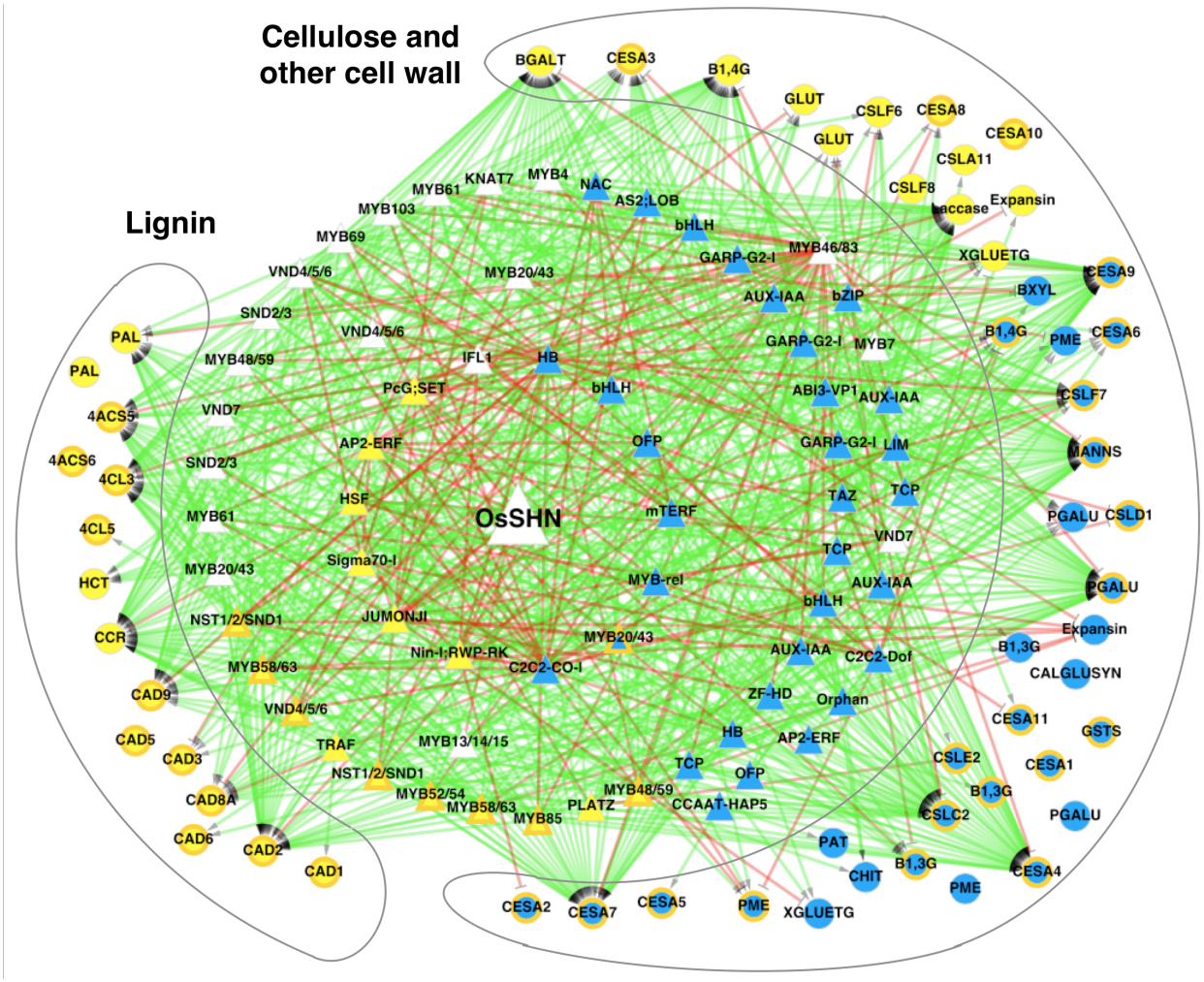


Figure 5.4: Coexpression network analysis and model of cell wall synthesis in rice.
Coexpression network connects TFs (triangles) to pathway gene targets (circles) through directed edges, and TFs to each other through undirected edges. Positive and negative correlation edges are colored green and red, respectively. Nodes are labeled with the gene name/family, and colored based on the direction of regulation in response to SHN expression (microarray data supplemented by qRT-PCR) with blue for up-regulation and yellow for down-regulation. Genes tested for differential expression using qRT-PCR are denoted with thick orange borders. The TFs in the outer ring are all positively coexpressed with each other, and hence, all the green edges connecting them to each other have been removed for clarity. See Table 5.S1 for IDs, names and annotations of all the genes in the network.

In this network, OsSHN was strongly connected to cell wall-associated TFs, which were amply connected to the biosynthetic genes. Overlaying the differential expression of the genes (from microarray and qRT-PCR) onto the nodes of the network showed that almost all the gene regulation observed due to the expression of Arabidopsis SHN in rice was explained by the signs of the rice network coexpression edges: positively correlated gene pairs were regulated in the same direction (both up- or both down-regulated), and negatively correlated gene pairs were regulated in opposite directions. OsSHN was directly connected to the homolog of VND6 through a negative edge, supporting the down-regulation of the VND6 homolog in the expression study. Here, for simplicity in transferring functional information between species, the rice TF genes are referred to by the names of Arabidopsis homologs (Zhong & Ye, 2009). VND6 is well connected to several other NAC (NST1/2/SND1) and MYB (MYB46/83, MYB52/54, MYB85 and MYB58/63) TFs, and to the lignin biosynthetic genes through positive edges, vindicating the down-regulation of this homologous transcriptional machinery by AtSHN leading to the repression of the lignin biosynthetic pathway. As expected, this large-scale repression also causes the down-regulation of several other cell wall pathway genes. On the other hand, among the putative cell wall TFs known from Arabidopsis, OsSHN is also directly positively connected to the MYB20/43 homolog, an interaction, again, supporting the up-regulation of this gene in response to AtSHN expression. This gene, in turn, is positively correlated with cellulose/cell wall genes that are also found to be up-regulated. Moreover, OsSHN is positively correlated with several other TFs (up-regulated by SHN expression), which are positively correlated with many cell wall genes found to be up-regulated in the expression studies including seven *CesA* genes and four CSL genes up-regulated by SHN expression.

Based on these observations, we hypothesize that OsSHN has a native association with cell wall regulatory and biosynthetic pathways, with an ability to coordinately regulate the lignin and cellulose pathways by shutting down the main switches (NACs), and intervening by directly regulating downstream MYBs: repressing MYBs specific to lignin biosynthesis (in addition to release of NAC activation), and activating MYBs and other TFs specific to cellulose/other cell wall biosynthesis.

5.3.3. AtSHN Directly Binds to the Promoters of NAC and MYB Genes

While independently seeking evidence for transcriptional regulation of NAC and MYB TF genes by SHN as observed in the coexpression expression analysis, the promoter regions of these genes were found to contain GCC-box motifs '[AG]CCGNC', known to be bound by AP2-ERF TFs (Ohme-Takagi & Shinshi, 1995) (Fig. 5.5), suggesting that AtSHN could regulate these TFs by direct binding. Hence, based on the coexpression network (Fig. 5.4), confirmed gene expression changes (using qRT-PCR; Fig. 5.3), and promoter analysis (Fig. 5.5), the NAC switches – VND6 (Os06g01480) and SND1/NST1/2 (Os08g02300, Os06g04090) – and three downstream MYBs – MYB20/43 (Os02g49986) and MYB58/63 (Os04g50770, Os02g46780) – were predicted to be direct targets of AtSHN. All these TFs have been shown to have roles in regulation of cell wall biosynthesis in Arabidopsis (Zhong & Ye, 2009).

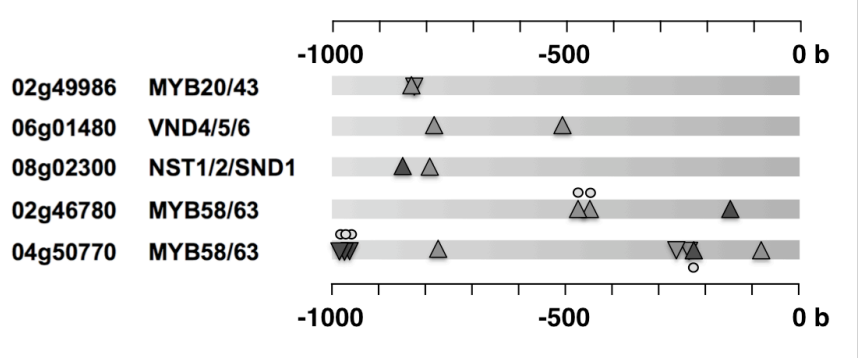


Figure 5.5: Locations of GCC-box motif '[AG]CCGNC' in the 1 Kb upstream sequences of the SHN regulated TF genes.
Upright and inverted triangles represent + and – strands, respectively. Dark triangles signify the presence of GCC-core 'GCCGCC'. Small circles above or below a triangle represent another overlapping GCC-box motif in the + or – strand, respectively.

To examine this hypothesis, we used recombinant 6His-AtSHN fusion protein for protein-DNA binding studies. Gel mobility shift assays showed binding between affinity-purified recombinant AtSHN protein and promoter regions of putative rice secondary wall TFs SND1/NST1/2, VND6, MYB58/63 and MYB20/43 (data not shown). Promoter analysis and binding assay, along with the observed gene expression changes, thus suggested that AtSHN could regulate these NAC and MYB TFs, the effect of which cascades into the biosynthetic pathways.

5.4. Discussion

Cellulose is the most abundant biopolymer and renewable energy resource, which can be processed into bioethanol as liquid fuel. In spite of its significance, very little is known about the regulation of cellulose biosynthesis and its co-regulation with the synthesis of other cell-wall biopolymers. We present a systems-level analysis on the complex regulation of cellulose and lignin biosynthesis, showing an unprecedented significant increase (1/3rd) in cellulose and decrease in lignin using the grass/crop model rice. These results offer novel ways for engineering non-food grasses and crop wastes for the production of lignocellulosic feedstocks that can be efficiently processed into biofuels. Expression of SHN leads to coordinate regulation of multiple steps in the pathways for monolignol and cellulose biosynthesis. Differential expression of genes was observed in the microarray, and was confirmed using qRT-PCR analysis using gene-specific primers of multi-gene family members at two different developmental stages of the plant (Fig. 5.2). SHN overexpressors had significantly repressed transcript levels of lignification genes such as CAD and 4CL family members (Fig. 5.2 *top*), and induced levels of cellulose synthase family and other cell wall genes (Fig. 5.2 *bottom*; see Fig. 5.S2) compared to wild-type (WT). It's noteworthy that OsCAD2, which is significantly repressed in SHN lines, is the rice ortholog of the maize CAD that maps to the *bm1* locus, knockout of which causes an 80% reduction in lignin (Tobias & Chow, 2005). Such large decrease in lignin content, either due to abolished CAD activity or transcriptional down-regulation as observed in SHN plants point to possibilities whereby grasses can be made more digestible. Likewise, in support of our hypothesis that SHN up-regulates hemicellulose and cellulose biosynthesis, several cellulose synthase-like GTs are up-regulated (Fig. 5.2 *bottom*), that are most likely involved in making hemicelluloses (Scheller & Ulvskov, 2010). In addition, three rice *CesA* genes (*OsCesA4, OsCesA7 and OsCesA9*) known to be involved in the synthesis of cellulose in the secondary cell walls (Tanaka et al, 2003) and responsible for the overall strength of the plant are up-regulated in SHN lines in both leaf and culm (Fig. 5.2).

Several other lines of evidence support the role of the SHN gene in coordinate regulation of cell wall synthesis pathways. First, the involvement of *AtSHN* in reduced lignin and increased cellulose levels correlates well with the regulation of NAC and MYB TFs (Fig 5.3) that control the activity of various branches of lignin and cellulose biosynthesis and the downstream

regulation of the biosynthetic genes (Fig. 5.1 and Fig 5.2). Second, the rice SHN phenotypic changes for lignin content are very similar to the effects reported for down-regulation of PAL, 4CL, CAD and target TF genes in *Arabidopsis* and other plants (Vanholme et al., 2008). Finally, Confocal microscopy and extensive biochemical analyses revealed evident qualitative differences in the lignin and cellulose levels of the SHN plants when compared with WT plants (data not shown).

Phenotype analysis and stem/leaf breaking force measurements of SHN lines showed that they have normal maturity and seed yield under greenhouse conditions, and unaltered strength of the stem/culm (data not shown). The tensile or bending strength of grass tissue, such as that of maize, has been shown to correlate with the cellulose content, whereas lignin is thought to play a role in resistance to compression (Dhugga, 2007). The increase in cellulose in secondary walls of SHN lines also probably offsets any reduction in mechanical strength due to reduced lignin. Such a compensatory increase in cellulose with reduced lignin has also been described by inhibition of a 4CL gene in aspen trees, which also exhibited better growth (Hu et al, 1999). Maize brown midrib mutants have also been shown to harbor reduced lignin and increased hemicellulose (with no change in cellulose content; (Vermerris et al, 2010)). On the other hand, the rice *brittle culm1* mutant and maize *brittle stalk2* mutant, which have reduced cellulose content, have been shown to contain more lignin (Ching et al, 2006; Li et al, 2003; Sindhu et al, 2007). These studies, along with ours presented here, provide evidence for complex interdependent regulation of the different cell wall pathways. However, we provide novel evidence for a TF – SHN – coordinating such a compensatory regulatory mechanism.

Coexpression network analysis was performed as an independent route to validating the function of SHN. Using a large gene expression compendium in rice, correlations between the expression profiles of all the cell wall regulatory and biosynthetic genes along with *OsSHN* were calculated. This was used as a predictive tool to first assess the expected associations between TFs and their targets, as in the case of NAC and MYB TFs and their putative lignin and other cell wall biosynthetic targets. Second, it was used to discover novel genes that might have a role in a pathway/process of interest using guilt-by-association, as in the case of *OsSHN*. And, finally, in conjunction with an independent gene expression dataset (here AtSHN expression microarray), it

was used to demarcate positive and negative interactions and propose a regulatory model for genes for interest. This analysis shows that *OsSHN*, the rice homolog of *AtSHN*, has a strong association with the cell wall regulatory and biosynthetic machinery in rice. The nature of association between the cell wall TFs and the biosynthetic genes is strongly positive as expected. The intriguing finding is the differential association of OsSHN with the TFs – via a negative connection to VND6 and a positive connection to MYB20/43 (and several other TFs), which are positively correlated with lignin and cellulose/other-cell-wall biosynthetic genes, respectively. In addition, most of the coexpression associations in the rice network agree with, and hence, corroborate the expression changes of the TFs and biosynthetic genes in response to AtSHN expression. Finally, AtSHN expressed in rice functions in the context of rice genes – established by the coexpression network – to perform its functions, which is the context in which OsSHN is expected to function. Taken together, these point to the underlying molecular mechanism by which SHN, in general, is able to achieve coordinate regulation of the cell wall pathways.

To further pursue these evidences, we were interested in finding if there were any clues for potential SHN regulation of the NAC and MYB TFs. We performed *de novo* motif discovery on the upstream regions of all the TFs in the network. However, the analysis showed no significant motifs, which we reasoned to be because SHN could actually bind to and regulate just a few major TFs that could then regulate other TFs and biosynthetic genes. Therefore, we restricted ourselves to the few TF candidates that showed verified gene expression changes and had homologs in Arabidopsis associated with the secondary wall biosynthetic pathways and we searched the upstream regions of these TF genes for the presence of GCC-box motif, a putative binding site of AP2-ERF TFs (Ohme-Takagi & Shinshi, 1995). Identification of several GCC-box motifs in these sequences (Fig. 5.5) motivated us to postulate that SHN could directly bind to and regulate these TFs, which we confirmed using mobility-shift binding assay (data not shown).

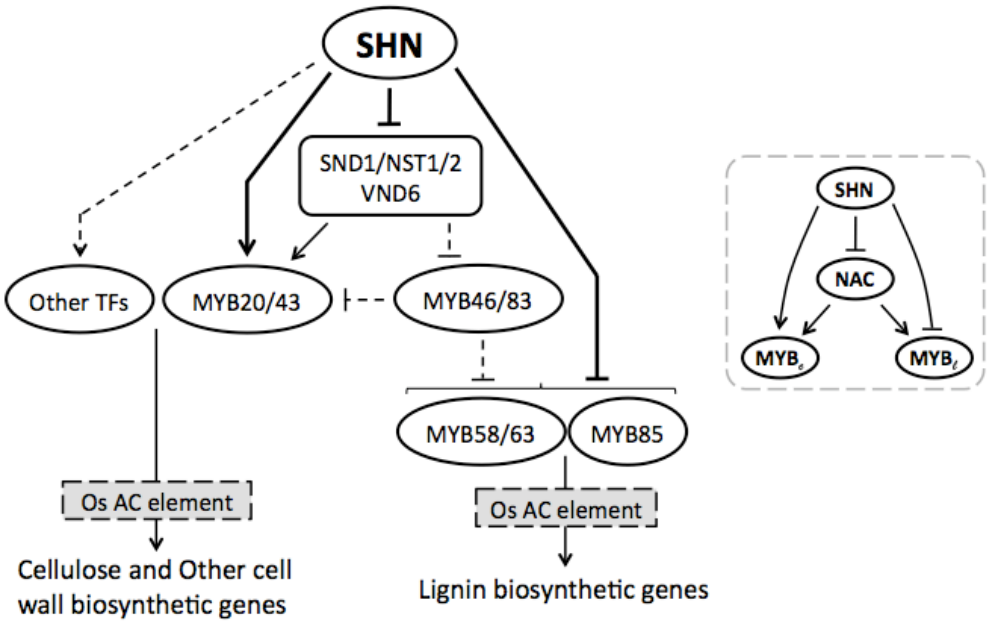


Figure 5.6: Hypothetical model of transcriptional regulation of cell wall biosynthesis in rice.
Lines with a pointed arrowhead signify activation and those with a tee signify repression. Dashed lines are hypothesized interactions based on the coexpression network and gene expression changes. Thick lines emanating from SHN represent confirmed interaction of SHN to the upstream regions of the TFs. *Inset*: NAC represents the NAC main switches, and MYBc and MYBl represent downstream MYB TFs hypothesized to be specific to cellulose/other cell wall genes and lignin genes, respectively.

The hypothetical model in Figure 5.6 summarizes our findings on the transcriptional regulation of cell wall biosynthesis in rice. SHN represses the NAC TFs, SND1/NST1/2 and VND6, which are known to be the main switches of cell wall biosynthesis. But, SHN also directly represses the MYB TFs MYB58/63 and MYB85, and directly activates MYB20/43. Due to these three activities, SHN bypasses the main switches (NACs) and selectively up- and down-regulates downstream TFs (MYBs) specific to cellulose (and other cell wall) and lignin biosynthetic genes, respectively. As depicted in the inset of Figure 6B, an interesting feature of this mechanism is its resemblance, in architecture, to a coupled system of feed-forward loops (Mangan & Alon, 2003), one incoherent (type 4; *left*) and other coherent (type 2; *right*); the relevance of this feature in a dynamical sense remains to be determined.

We also hypothesize that there might be other TFs involved in mediating SHN up-regulation of cellulose and other cell wall genes, candidates for which have been identified as those up-regulated in response to *SHN* expression and positively correlated with the up-regulated cell wall genes. In this context, as an aside, we posit from the coexpression network an aspect about the

regulation of MYB46/83, a gene not necessarily regulated by SHN, but shown to have a role in regulation of cell wall biosynthesis (McCarthy et al, 2009): in Arabidopsis it is activated by the NAC main switches whereupon it activates the rest of the downstream TFs (MYBs and potentially other TFs). However, since MYB46/83 is negatively correlated to both the NAC switches and the downstream MYBs in rice, we hypothesize that MYB46/83 mediates NAC-activation of downstream MYBs through a double negative route (Fig. 5.6).

Since several insights about cell wall biosynthesis have been gleaned in Arabidopsis it is also important to put our findings in rice in that context and deliberate the probable role of AtSHN in Arabidopsis. To gain some understanding about AtSHN function in relation to secondary cell wall biosynthesis, we re-analyzed the gene expression profiles of Arabidopsis AtSHN (*WIN1*) overexpression lines compared to WT controls from Broun et al. (2004). Since the design lacks replication, we quantified the relative expression levels of ~390 (present in the 8K Affymetrix Arabidopsis genome array) cell-wall-related genes (out of ~930 total in the genome) in a 'strong' WIN1 overexpressor compared to WT (see Table 5.S2). We observed here that quite a few lignin (including PAL2, PAL3 and 4CL3) and cellulose/hemicellulose biosynthetic genes (including CESA2, CSLB01, CSLB02, CSLB03, CSLB04, CSLC12, and CSLG3) were up-regulated. On the other hand, two cellulose genes (CESA1 and CSLA03) and two lignin genes (CCR and CCoA-OMT) were down-regulated. Moreover, along with the biosynthetic genes, the NAC TF NST1 (functional paralog of SND1) is also up-regulated.

Based on the above observations, overexpression of AtSHN in Arabidopsis seems to cause a nominal up-regulation (along with certain amount of down-regulation) of both lignin and cellulose/hemicellulose biosynthesis, and therefore probably does not cause any large absolute or relative change in the amounts of lignin and cellulose. In support of this notion, simple sugar measurements made in the Kannangara et al. (2007) paper and our observations of Arabidopsis cross-sections for lignin quantification, respectively, indicate that there are no significant changes in either cellulose or lignin in Arabidopsis AtSHN lines compared to WT.

Nevertheless, we propose that AtSHN does have an association with secondary cell wall biosynthesis in Arabidopsis as far as transcriptional regulation of the regulatory and biosynthetic

genes is concerned. This includes AtSHN's ability to both up- and down-regulate the involved genes. Therefore, an overall similar association with secondary cell wall regulation and biosynthesis is conserved across Arabidopsis and rice, with different details: we show that AtSHN in rice causes an inverse regulation of the pathways by differently regulating various TFs of the lignin and cellulose biosynthetic pathways.

Furthermore, to explore the expression pattern of OsSHN in different rice organs and tissues, we used the rice eFP browser (Winter et al, 2007) to identify that the gene is expressed in the inflorescence stages P3, P4, P5 and P6, with the strongest expression at the P5 stage (Fig. 5.S3). Using the rice expression atlas (Jiao et al, 2009) resource, we observed that OsSHN is expressed in the coleoptile (0hr) and fresh whole leaf and to a lesser extent in epiblast (12hr) and seedling blade (Table 5.S3). These cell types represent young growing tissue where lower lignin deposition is expected, consistent with SHN's proposed role in up-regulating cellulose synthesis and down-regulating lignin synthesis pathways in rice.

The SHN master regulator orchestrates coordinated regulation of the cellulose and lignin pathways to provide enhanced cellulose and decreased lignin deposition. The two other functions ascribed to SHN/WIN are in cuticle (cuticular wax and cutin) formation (Aharoni et al, 2004; Broun et al, 2004; Kannangara et al, 2007). Most strikingly, all these processes could have evolved in organismal organization simultaneously when land plants emerged, to give them a protective cover as well as strength to remain erect and transport water upwards. Coordination of these processes was probably maintained by the master regulatory functions of the *SHN* gene family. Altogether, SHN regulates the accumulation of cellulose, lignin and cutin, the top three plant biomass polymers, and can help in improving plant feedstock for these components.

Lignocellulose is the major biomass feedstock useful for converting into biofuels such as ethanol. The activity of SHN in the grass model rice shows that it has a potential role in engineering the cellulose-lignin composition and content in grasses or other suitable biomass producers. The SHN master regulator can also be used as a tool, to express in plants of interest and unravel the regulatory pathway in cellulose and lignin biosynthesis. The Arabidopsis-rice homologous transcription factors and biosynthetic genes described here, and the genetic model

of their interaction, can serve as a dicot-monocot conserved model to understand the coordinate regulation of cellulose and lignin biosynthesis in a number of other plant species.

5.5. Methods

5.5.1. Gene Expression Analysis

Total RNA was isolated from the rice leaf and culm tissue of WT and SHN lines using the RNeasy plant kit (Qiagen, USA), RNA quantity/quality measured by the Agilent 2100 Bioanalyzer (Agilent Technolgies, USA). For each sample 4 µg total RNA was used to generate first-strand cDNA with a T7-Oligo(dT) primer. Following second-strand synthesis, in vitro transcription was performed using the GeneChip® IVT Labeling Kit. The preparation and processing of labelled and fragmented cRNA targets, as well as hybridization to rice Affymetrix GeneChips, washing, staining, and scanning were carried out according to manufacturer's instructions (http://www.affymetrix.com). The RNA samples used for the microarray experiments were also used to synthesize cDNA templates for qRT-PCR analysis.

5.5.2. Reannotation of Rice Genechip Probe-Gene Mapping

A high-quality custom chip definition file (CDF) was built for the rice GeneChip array by uniquely mapping 442,810 probe sequences (http://www.affymetrix.com/analysis/downloads/data/) to 35,161 rice gene-based probesets in the following manner: (i) probes that have perfect sequence identity with a single target gene were selected, (ii) probes mapping to reverse complements of genes were annotated separately as antisense probes (not used in the above counts), and finally, (iii) probes were grouped into probe sets, each corresponding to a single gene, and probe sets with at least 3 probes were retained (>98% probe sets have >=5 probes). Note that these stringent criteria used to construct the CDF make it possible to reliably measure expression values of members of multigene families (free from cross-hybridization between paralogs showing high sequence similarity) and to get around 'one gene to multiple probesets' ambiguities.

5.5.3. Analysis of Differential Gene Expression

Raw data from the *SHN* overexpression experiment were background corrected, normalized and summarized according to the custom CDF using RMA (Gentleman et al, 2004; Ihaka &

Gentleman, 1996; Irizarry et al, 2003), followed by non-specific filtering of genes that do not have enough variation (interquartile range (IQR) across samples < IQR_{median}) to allow reliable detection of differential expression. A linear model was then used to detect differential expression of the remaining genes (Smyth, 2004). The *p*-values from the moderated *t*-tests were converted to *q*-values to correct for multiple hypothesis testing (Storey & Tibshirani, 2003), and genes with *q*-value <0.1 were declared as differentially expressed in response to *AtSHN* expression. WIN1 overepxression data was obtained from NCBI GEO accession GSE1071 and raw data was similarly pre-preprocessed using RMA based on a custom CDF for the Arabidopsis Genome array obtained from http://brainarray.mbni.med.umich.edu/Brainarray/ (Dai et al, 2005).

5.5.4. Curation of Lignin and Cellulose Biosynthetic Genes and Putative Regulators

Rice genes involved in cell wall biosynthesis and regulation were identified as homologs of *Arabidopsis* cell wall-related genes (Remm et al, 2001; Yokoyama & Nishitani, 2004; Zhong & Ye, 2009) and from direct annotations in the Rice genome annotation database (Ouyang et al, 2007). A further 153 transcription factors that were regulated in the SHN expression microarray (*see below*) were added to the list of putative regulators.

5.5.5. Coexpression Network Analysis

29 publicly available Affymetrix rice GeneChip gene expression datasets (414 samples; 150 groups after gathering biological replicate samples into single groups) were collected from NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/) (Barrett et al, 2009) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) (Parkinson et al, 2009), and the largest subset of experiments (10 datasets; 129 samples; 45 groups) with a similar biological context corresponding to studies of response to some environmental condition was used for coexpression analysis (see Table 5.S4).

Raw data were background corrected, normalized and summarized according to the custom CDF using justRMA (Irizarry et al, 2003), and expression values were averaged across replicates. Pearson correlations were first calculated between every pair of genes (Huttenhower et al, 2008) (see *Note* at the end of this subsection), which were then Fisher Z-transformed (David, 1949) and standardized to get coexpression scores (z_{cs}) with a N(0,1) distribution. This formulation was

robust and highly interpretable as deviations from the expected value, and even by level of significance where $|z_{cs}|$ values greater than 1.645, 1.96 and 2.58 correspond to 10%, 5% and 1% extremes of the distribution of z_{cs} scores.

TFs of interest were connected to each other when they had a strong correlation ($|z_{cs}| > 1.645$). TFs were connected to pathway genes based on a more rigorous procedure: For each set of 'pathway' genes, 'lignin' and 'other cell wall', a pathway correlation matrix was first created taking the pathway genes along the columns, and the pathway genes plus the putative TFs along the rows. Here, cell (i, j) contained the coexpression score $z_{cs}(i, j)$ between genes i and j. This was same as the adjacency matrix of the pathway genes, only with extra rows of TFs. Thus, each row i contained the vector of z_{cs}'s of gene i to all the pathway genes, measuring its 'association' with the entire pathway. Pearson correlations were then calculated between rows, and TF-pathway-gene pairs with absolute correlation >0.8 were selected. This correlation measures how well the genes agree with the regulatory program of the pathway. TF-TF edges were left undirected while TF-pathway-gene edges were directed from TF to putative target. Among the pathway genes, only those regulated by SHN (from the expression studies) were included in the final network. The network was visualized using Cytoscape (Shannon et al, 2003).

Note: There are popular methods to derive regulatory networks from gene expression data based on calculation of mutual information (MI) – for example, ARACNE (Basso et al, 2005) and CLR (Faith et al, 2007) – that perform better than simple correlation-based methods. But, these methods require very large amounts of gene expression data to contain expression of the genes across a large dynamic range to calculate MI reliably. Hence, with relatively less data in rice, especially when considering those similar in biological context, using MI-based methods would not be possible. Then, for measuring simple correlations, the Spearman rank correlation metric is a good choice. But, given we have carefully chosen datasets similar in biological context, identical in experimental platform, and resolved ambiguity in hybridization using a redefinition of probe-gene mapping in the array, we sought to using a metric more sensitive to the actual expression values, like the Pearson correlation coefficient, rather than one that works on the relative ranks of the values, like the Spearman rank correlation coefficient.

5.5.6. Promoter Analysis

For promoter analysis in rice, FIRE (Elemento et al, 2007) was used to discover motifs specific to the cell wall pathway genes by comparing the motif content of 1 Kb upstream sequences of these genes to that of the rest of the genome, followed by comparison to known cis-elements (Crooks et al, 2004; Higo et al, 1999; Mahony & Benos, 2007). This *de novo* approach was taken since cis-regulatory motifs could diverge quickly across species making them hard to find simply by searching. A Perl script was used to search for GCC-box motifs '[AG]CCGNC' in the 1 Kb upstream sequences of SHN-regulated TFs.

5.6. References

Aharoni A, Dixit S, Jetter R, Thoenes E, van Arkel G, Pereira A (2004) The SHINE clade of AP2 domain transcription factors activates wax biosynthesis, alters cuticle properties, and confers drought tolerance when overexpressed in Arabidopsis. *Plant Cell* **16:** 2463-2480

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* **37:** D885-890

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet* **37:** 382-390

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

Broun P, Poindexter P, Osborne E, Jiang CZ, Riechmann JL (2004) WIN1, a transcriptional activator of epidermal wax accumulation in Arabidopsis. *Proc Natl Acad Sci U S A* **101:** 4706-4711

Ching A, Dhugga KS, Appenzeller L, Meeley R, Bourett TM, Howard RJ, Rafalski A (2006) Brittle stalk 2 encodes a putative glycosylphosphatidylinositol-anchored protein that affects

mechanical strength of maize tissues by altering the composition and structure of secondary cell walls. *Planta* **224:** 1174-1184

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* **14:** 1188-1190

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33:** e175

David FN (1949) The moments of the Z and F distributions. *Biometrika* **36:** 394–403

Dhugga KS (2007) Maize Biomass Yield and Composition for Biofuels. *Crop Sci* **47:** 2211–2227

Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* **28:** 337-350

Faith J, Hayete B, Thaden J, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins J, Gardner T (2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biol* **5:** e8

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge YC, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G et al (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5:** -

Haigler CH, Ivanova-Datcheva M, Hogan PS, Salnikov VV, Hwang S, Martin K, Delmer DP (2001) Carbon partitioning to cellulose synthesis. *Plant Mol Biol* **47:** 29-51

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

Hu WJ, Harding SA, Lung J, Popko JL, Ralph J, Stokke DD, Tsai CJ, Chiang VL (1999) Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. *Nat Biotechnol* **17:** 808-812

Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG (2008) The Sleipnir library for computational functional genomics. *Bioinformatics* **24:** 1559-1561

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. *J Comput Graph Stat* **5:** 299-314

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003) Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Research* **31:** -

Jakob K, Zhou FS, Paterson A (2009) Genetic improvement of C4 grasses as cellulosic biofuel feedstocks. *In Vitro Cell Dev-Pl* **45:** 291-305

Jiao Y, Tausta SL, Gandotra N, Sun N, Liu T, Clay NK, Ceserani T, Chen M, Ma L, Holford M, Zhang HY, Zhao H, Deng XW, Nelson T (2009) A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nat Genet* **41:** 258-263

Kannangara R, Branigan C, Liu Y, Penfield T, Rao V, Mouille G, Hofte H, Pauly M, Riechmann JL, Broun P (2007) The transcription factor WIN1/SHN1 regulates Cutin biosynthesis in Arabidopsis thaliana. *Plant Cell* **19:** 1278-1294

Karaba A (2007) Improvement of water use efficiency in rice and tomato using Arabidopsis wax biosynthetic genes and transcription factors., Wageningen University, Netherlands.,

Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J, Mimura T, Fukuda H, Demura T (2005) Transcription switches for protoxylem and metaxylem vessel formation. *Genes Dev* **19:** 1855-1860

Li X, Weng JK, Chapple C (2008) Improvement of biomass through lignin modification. *Plant J* **54:** 569-581

Li X, Yang Y, Yao J, Chen G, Zhang Q, Wu C (2009) FLEXIBLE CULM 1 encoding a cinnamyl-alcohol dehydrogenase controls culm mechanical strength in rice. *Plant Mol Biol* **69:** 685-697

Li Y, Qian Q, Zhou Y, Yan M, Sun L, Zhang M, Fu Z, Wang Y, Han B, Pang X, Chen M, Li J (2003) BRITTLE CULM1, which encodes a COBRA-like protein, affects the mechanical properties of rice plants. *Plant Cell* **15:** 2020-2031

Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* **35:** W253-258

Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. *Proc Natl Acad Sci U S A* **100:** 11980-11985

McCarthy RL, Zhong R, Ye ZH (2009) MYB83 is a direct target of SND1 and acts redundantly with MYB46 in the regulation of secondary cell wall biosynthesis in Arabidopsis. *Plant Cell Physiol* **50:** 1950-1964

Mele G, Ori N, Sato Y, Hake S (2003) The knotted1-like homeobox gene BREVIPEDICELLUS regulates cell differentiation by modulating metabolic pathways. *Genes Dev* **17:** 2088-2093

Ohme-Takagi M, Shinshi H (1995) Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell* **7:** 173-182

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res* **35:** D883-887

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ et al (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* **37:** D868-872

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314:** 1041-1052

Scheller HV, Ulvskov P (2010) Hemicelluloses. *Annu Rev Plant Biol* **61:** 263-289

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498-2504

Sindhu A, Langewisch T, Olek A, Multani DS, McCann MC, Vermerris W, Carpita NC, Johal G (2007) Maize Brittle stalk2 encodes a COBRA-like protein expressed in early organ development but required for tissue flexibility at maturity. *Plant Physiol* **145:** 1444-1459

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

Somerville C (2006) Cellulose synthesis in higher plants. *Annu Rev Cell Dev Biol* **22:** 53-78

Sticklen M (2006) Plant genetic engineering to improve biomass characteristics for biofuels. *Curr Opin Biotechnol* **17:** 315-319

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100:** 9440-9445

Taketa S, Amano S, Tsujino Y, Sato T, Saisho D, Kakeda K, Nomura M, Suzuki T, Matsumoto T, Sato K, Kanamori H, Kawasaki S, Takeda K (2008) Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. *Proc Natl Acad Sci U S A* **105:** 4062-4067

Tanaka K, Murata K, Yamazaki M, Onosato K, Miyao A, Hirochika H (2003) Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. *Plant Physiol* **133:** 73-83

Tobias CM, Chow EK (2005) Structure of the cinnamyl-alcohol dehydrogenase gene family in rice and promoter activity of a member associated with lignification. *Planta* **220:** 678-688

Vanholme R, Morreel K, Ralph J, Boerjan W (2008) Lignin engineering. *Curr Opin Plant Biol* **11:** 278-285

Vermerris W, Sherman DM, McIntyre LM (2010) Phenotypic plasticity in cell walls of maize brown midrib mutants is limited by lignin composition. *J Exp Bot* **61:** 2479-2490

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ (2007) An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. *PLoS ONE* **2:** e718

Yokoyama R, Nishitani K (2004) Genomic basis for cell-wall diversity in plants. A comparative approach to gene families in rice and Arabidopsis. *Plant Cell Physiol* **45:** 1111-1121

Zhong R, Demura T, Ye ZH (2006) SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. *Plant Cell* **18:** 3158-3170

Zhong R, Ye ZH (2009) Transcriptional regulation of lignin biosynthesis. *Plant Signal Behav* **4:** 1028-1034

Zhou J, Lee C, Zhong R, Ye ZH (2009) MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. *Plant Cell* **21:** 248-266

6. Mutant Resources in Rice for Functional Genomics of the Grasses

Arjun Krishnan, Emmanuel Guiderdoni, Gynheung An, Yue-ie C. Hsing, Chang-deok Han, Myung Chul Lee, Su-May Yu, Narayana Upadhyaya, Srinivasan Ramachandran, Qifa Zhang, Venkatesan Sundaresan, Hirohiko Hirochika, Hei Leung, Andy Pereira

Plant Physiology (2009) 149(1): 165-1670.

Used with permission of American Society of Plant Biologists.

6.1. Introduction

Rice is the reference genome for the grasses including cereals. The complete genome sequence lays the foundation for comparative genomics to the other grasses based on genome structure and individual gene function (Devos, 2005; International Rice Genome Sequencing Project, 2005). The basic complement of monocot genes in rice can be examined by functional genomics studies, because of the many advantages of rice as a system for genetic analysis as well as the worldwide development of resources.

The analysis of mutants by forward and reverse genetics approaches is an effective way to study gene function. Knockout (KO) mutations, which abolish gene expression and display a phenotype, provide a direct causal relationship between the gene sequence and its biological function. However, not all gene mutations display a KO mutant phenotype, primarily due to gene redundancy since plant genomes have been shown to have large segmental genomic duplications as well as tandem duplications of gene families (Sterck et al, 2007; Yu et al, 2005). In many cases the redundancy is partial or unequal due to overlap in expression of duplicated genes (Briggs et al, 2006), or the gene activity is required only under some specific conditions such as biotic/abiotic stresses where the mutant phenotype can be observed.

The use of molecular “tags” or DNA insertions such as transposons or T-DNA are favored for mutations as their genome positions can be easily monitored to determine the correlations between tagged genes and phenotypes. The limitations in identifying gene functions by KO mutations alone, are resolved by employing heterologous DNA insertions with engineered

properties to monitor the expression of tagged genes using entrapment vectors, or to alter the expression of tagged genes using activation tagging (Pereira, 2000). The wide variety of mutants required for genome-wide functional genomics in the grasses could be obtained by developing large-scale mutant resources in rice for community use.

The International Rice Functional Genomics Consortium (IRFGC), combined with many national programs set a goal to generate mutant resources towards discovering the function of all rice genes, primarily through reverse genetics approaches (Hirochika et al, 2004). This resource update describes the generation of over 200,000 insertion flanking sequence tags (FST), which tag two-third of the predicted protein coding genes, with half of the protein coding genes estimated to have knockout mutations. The insertion sequences comprise the endogenous *Tos17* retrotransposon, modified maize *Ds* and *dSpm* elements, and *Agrobacterium* T-DNA. An analysis of the genome distribution of these three types of insertions in the genome shows an insertion bias, with *Tos17* having the highest insertions in exons. The heterologous transposon and T-DNA inserts engineered to function as enhancer trap, gene trap and activation tags have been demonstrated to be very useful in gene function identification, In addition, chemical and physical mutagen derived mutant populations have been developed that are available for TILLING and other high-throughput screens. The extensive number and variety of mutant resources described here for rice are very amenable for dissecting the functions of genes of interest in other grasses.

6.2. Development of rice mutant resources

6.2.1. Insertional mutants

With the sequencing of plant genomes it was recognized that insertional mutants that are indexed by their insertion position in the genome would be very suitable for systematic analysis of annotated genes by reverse genetics (Parinov & Sundaresan, 2000). Extensive mutant collections defined by insertion positions are now available in Arabidopsis databases (http://signal.salk.edu/Source/AtTOME_Data_Source.html), which comprise a total of 379,674 inserts tagging 30,280 out of the predicted 33,003 genes.

In rice, the two component maize transposons *Ac-Ds* (Chin et al, 1999; Greco et al, 2003; Kolesnik et al, 2004; Upadhyaya et al, 2002) and *En*/*Spm-dSpm* (Greco et al, 2004; Kumar et al,

2005) have been well characterized for their activity. The early studies revealed some problems in transposon silencing and uncontrolled activity, but since then good genetic selection systems have been developed to select for transpositional activity that is usable for large-scale mutagenesis (Hirochika et al, 2004; Zhu et al, 2007). In addition, the endogenous rice *Tos17* retrotransposon is active in specific genotypes and conditions and is an effective insertion mutagen in the rice genome (Miyao et al, 2003). The development of efficient protocols for rice transformation has helped in the generation of a large number of transgenic rice plants bearing low copy T-DNA insertions (Jeon et al, 2000; Sallaud et al, 2003).

In the native species (e.g. in maize), transposons have been very useful for KO or loss-of-function mutagenesis. However, the engineering of transposon and T-DNA constructs offers immense flexibility in fashioning the insertion sequences to detect adjacent gene expression or activate the expression of adjacent genes by activation tagging resulting in gain-of-function mutations. These modified insertions can contribute to gene function discovery of redundant genes and those having lethal mutant effects.

Gene entrapment*:* To facilitate the analysis of genes based on their expression patterns, gene trap (GT) and enhancer trap (ET) constructs have been designed that carry a reporter gene and can display the expression pattern of an adjacent trapped gene (Sundaresan et al, 1995). The reporter gene pattern in ET inserts reflects the adjacent plant gene enhancer activity, and in GT inserts the adjacent gene promoter activity. ET and GT constructs have been used in both T-DNA and *Ac-Ds* transposons in rice yielding interesting gene expression patterns and entrapped genes, which support their widespread generation and use for complementing KO mutagenesis (An et al, 2005; Hirochika et al, 2004).

Activation tagging*:* A T-DNA activation tag (AT) population was developed using a vector with CaMV 35S enhancer tetramer (Jeong et al, 2002), and FSTs generated to facilitate reverse genetics screens (Jeong et al, 2006). Recently an *Ac-Ds* AT system has also been developed (Qu et al, 2008) using convenient markers for selection of multiple transposants from a few starter transformed lines. In both these AT systems, activation of adjacent genes is observed, albeit

52.7% of the T-DNA lines and 20.8% of the *Ds* tags activate adjacent genes, which can be as far away as 10 kb from the AT enhancer.

Institution	Genotype	Mutagen	Mutated loci	FSTs/ screen	FST Lines availability[1]	Database website	Contact
CIRAD-INRA-IRD-CNRS, Génoplante, FR	Nipponbare	T-DNA ET *Tos17*	45,000 100,000	14,137 13,745	13,600 768 (March 2009)	http://urgi.versailles.inra.fr/OryzaTagLine/	E. Guiderdoni guiderdoni@cirad.fr
CSIRO Plant Industry, AU	Nipponbare	*Ac-Ds* GT/ET	16,000	611	~ 50% lines no seed	http://www.pi.csiro.au/fgrttpub/	N.M. Upadhyaya narayana.upadhyaya@csiro.au
EU-OSTID, EU	Nipponbare	*Ac-Ds* ET	25,000	1380	1300	http://orygenesdb.cirad.fr/	E. Guiderdoni guiderdoni@cirad.fr
IRRI, PH	IR64	Fast neutron γ-ray DEB, EMS	500,000	Deletion database: 400 genes		http://www.iris.irri.org/cgi bin/MutantHome.pl	H. Leung H.Leung@cgiar.org
Gyeongsang National Univ., KR	Dongjin Byeo	*Ac-Ds* GT	30,000	4820	4820	KRDD http://www.niab.go.kr/RDS/	C-D Han cdhan@nongae.gsnu.ac.kr
NIAS, JP	Nipponbare	Tos 17	500,000	34,844		http://tos.nias.affrc.go.jp	H. Hirochika hirohiko@nias.affrc.go.jp
NIAS, JP	Nipponbare	γ-ray ion beam	15000 M2 7000 M2	DNA pools			M. Nishimura nisimura@affrc.go.jp
POSTECH, KR	Dongjin, Hwayoung	T-DNA ET/AT *Tos17*	150,000 400,000	84,680	58,943	RISD http://an6.postech.ac.kr/pfg/	G. An genean@postech.ac.kr
Huazhong Agricultural Univ., CN	Zhonghua 11 Zhonghua 15 Nipponbare	T-DNA ET	113,262 14,197 1,101	16,158	26,000 Dec 2008	RMD http://rmd.ncpgr.cn/	Q. Zhang qifazh@mail.hzau.edu.cn
SIPP, CN	Zhonghua 11	T-DNA ET	97,500	8,840	8,840 FST +11,000 lines	http://ship.plantsignal.cn/home.do	F. Fu ship@sibs.ac.cn
Temasek Lifesciences, SG	Nipponbare	*Ac-Ds* GT	20,000	3500			R. Srinivasan sri@tll.org.sg
IPMB, Academia Sinica, TW	Tainung 67	T-DNA AT	30,000	18,382	31,000	TRIM http://trim.sinica.edu.tw	Y.C. Hsing bohsing@gate.sinica.ed.tw
University of California-Davis, US	Nipponbare	*Ac-Ds* GT *Spm*/*dSpm*	20,000	*Ds* 4,735 *dSpm* 9,469	4,630 9,036	http://www-plb.ucdavis.edu/Labs/sundar/	V. Sundaresan sundar@ucdavis.edu
University of California-Davis	Nipponbare	Na azide +MNU	6,000	TILLING screen		http://tilling.ucdavis.edu/	L. Comai lcomai@ucdavis.edu
Zhejiang University, CN	Nipponbare Zhonghua 11	T-DNA		1009	1009	http://www.genomics.zju.edu.cn/ricetdna	P. Wu clspwu@zju.edu.cn
Zhejiang University, CN	Kasalath SSBM	γ-ray EMS	40,000			http://www.genomics.zju.edu.cn	P. Wu clspwu@zju.edu.cn

Table 6.1: Mutant Resources, Contributors and Databases.

[1]Based on searching current project database, future plans, and subject to seed availability. Institution abbreviations: CIRAD, Centre de Coopération Internationale en Recherche Agronomique pour le Développement; CNRS, Centre National de la Recherche Scientifique; INRA, Institut National de la Recherche Agronomique; IPMB, Institute of Plant and Microbial Biology; IRD, Institut de Recherche pour le Développement; IRRI: International Rice Research Institute; NIAS, National Institute of Agrobiological Sciences; POSTECH, Pohang University of Science and Technology; SIPP, Shanghai Institute of Plant Physiology and Ecology.

6.2.2. Chemical and physical mutagenesis

Chemical agents such as ethyl methanesulphonate (EMS), nitrosomethylurea (NMU) and DB, or physical methods like fast neutron, γ-rays and ion beam irradiation can cause a high density of mutations which can saturate the genome (Hirochika et al, 2004). Mutant populations have been generated in rice, in which the point mutations can be screened by TILLING and larger deletions by PCR based screens (Till et al, 2007; Wu et al, 2005). The IR64 (Wu et al, 2005) and the Nipponbare (Till et al, 2007) populations as well as other unpublished populations also shown in Table 6.1, offer different backgrounds and mutation spectrum. The IR64 mutant collection comprises a total of 66,891 mutant lines in the M4 generation. Of these, about 15,000 are γ-ray-induced mutants, each carrying 30 to 40 deletions per genome, thus contributing to a conservative estimate of over 500,000 mutations in the collection (H. Leung, unpublished). Screening for mutations in such populations can be done using genome-wide chips or other high-throughput genotyping technologies. It is expected that these populations and/or the DNA pools will soon be publicly available for screening purposes.

6.3. Utility of mutant resources for functional genomics in rice

6.3.1. Forward and Reverse genetics in rice

The first rice genes identified by insertional mutagenesis were with *Tos17* in a forward genetics screen for viviparous mutants (Agrawal et al, 2001), and simultaneously in a reverse genetics screen for inserts in phytochrome A genes (Takano et al, 2001). With T-DNA, genes were identified by forward screens (Jung et al, 2003), by reverse genetics PCR-based screens for mutations in specific genes (Lee et al, 2003) as well as expression based GT screens (Kang et al, 2005). Likewise, the maize *Ac-Ds* transposon system also yielded tagged genes (Zhu et al, 2003; Zhu et al, 2004).

Since the complete genome sequence of rice became available, the generation of FST information of mutant populations has made the mutants more accessible to address biological questions. Table 6.1 shows the different mutant populations available and the FSTs that can be screened for inserts in genes of interest. Such queries can be made *in silico,* thus providing a convenient way to assess mutant populations around the world. The *Ds* and *dSpm* insertions are generated by transposition from a few starter-transformed lines and do not directly result from a regeneration process (Kolesnik et al, 2004; Park et al, 2007; Qu et al, 2008; Upadhyaya et al, 2006; van Enckevort et al, 2005). The FST resource of the endogenous *Tos17* retrotransposon are generated by regeneration process in Nipponbare (Miyao et al, 2003) which also accompanies *Agrobacterium* transformation of T-DNA yielding additional *Tos17* insertions (Piffanelli et al, 2007). The T-DNA insertions with FSTs in various genetic backgrounds comprise an extensive diverse resource (Chen et al, 2003; Hsing et al, 2007; Jeong et al, 2006; Sallaud et al, 2004; Zhang et al, 2006).

The generation of insertions accompanied by a regeneration phase such as for T-DNA and *Tos17* can result in a high frequency of untagged mutations in the background that can complicate genetic analysis of the mutants (H. Leung, E. Guiderdoni unpublished). To alleviate this problem, genetic segregation analysis and the use of multiple mutants of the gene would be useful. Alternatively, the use of transposon reversions that restore the wild-type phenotype is a convenient approach to prove gene-phenotype relationships.

6.3.2. Resources and databases for reverse genetics

To facilitate the identification of insertion mutations in genes using available FST information, a number of project database websites are available as shown in Table 1. In addition, functional genomics databases are available such as RiceGE/SIGnAL (http://signal.salk.edu/cgi-bin/RiceGE), OryGenesDB (http://orygenesdb.cirad.fr/) and Gramene (http://www.gramene.org/) where the FST information has been collated and mutants can found for inserts in genes of interest. These databases link rice genes to other grass genes, and thus direct functional queries to the rice mutant resources.

6.4. Properties of insertion mutants

We compiled 206,668 insertion FSTs from our contributing groups, which comprise 180,639 unique hits in the genome (Table 6.S1). The different insertion types (*Tos17*, T-DNA, *Ds*, *dSpm*) show differences in their specificity, with *Tos17* showing the highest proportion of insertions in exons (Fig. 6.1). A remarkably large proportion of all the inserts (62.5%) are in genic regions, including 5' and 3' regions, as described in Figure 6.1.

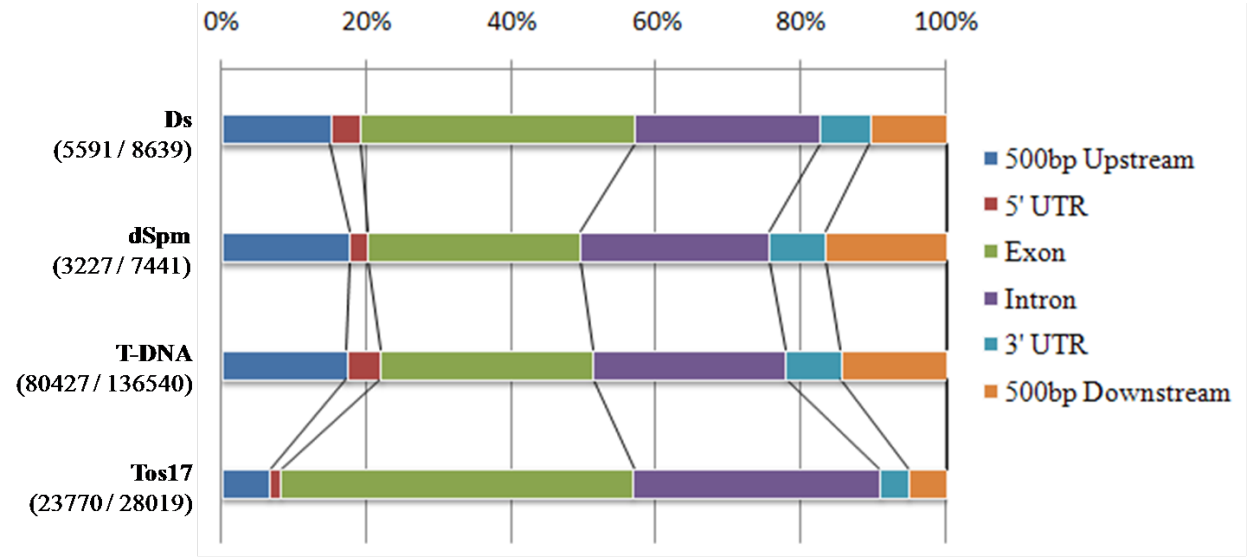


Figure 6.1: Distribution of insertion positions within genic regions in rice.
The three classes of insertion mutagens: endogenous *Tos17* retrotransposon, *Agrobacterium* T-DNA, and maize cut-and-paste transposons (*Ds* and *dSpm*), shown with their insertion positions in different parts of genes with color-code shown alongside. The numbers of individual insertion types in genes in relation to the total number of insertions are entered below the insertion name. Datasets from the following resources were used: *Ds* – CSIRO, OSTID, UCD and KRDD_GNU; *dSpm* – UCD; *T-DNA* – RIFGP_ZJU, RMD, SHIP, PFG, CIRAD and TRIM; *Tos17* – AFFRC_NIAS and CIRAD. 500bp Upstream and Downstream regions correspond to sequences upstream and downstream of the transcription unit (start site to site of termination). See Table 6.S1 for more information.

However, many genes have multiple different insertions, with a total of 32,459 genes containing inserts out of the total 56,985 (56.9%) nuclear genes with assigned locus IDs. Among the 41,753 predicted protein-coding rice genes, 28,545 (68.4%) have inserts in the genic region. Assuming that the most probable insertions to produce KO mutations would be those in exons, introns and the 5'UTR, the insertions were recalculated to be 21,239 (50.8%) in the protein coding genes (Table 6.S1). One of the major reasons for a low frequency of insertions in genes is the actual target size, with around 13,000 genes of 1 Kb size showing only around 35% bearing insertions (Fig. 6.2). And, as expected, the percentage of tagged genes increases more or less steady with

increase in target gene size. The insertion mutants found for the rice annotated genes, defined by the GOslim biological process (10,232 total) and molecular function (12,765) categories (Fig. 6.3 and 6.4) reveal an even distribution of >90% in total genic region and around 80% in the critical KO mutation target region. This shows that a high proportion of mutations in annotated genes would most probably cause KO mutants, while the frequencies in the unannotated genes is relatively lower. However, some genes annotated to be involved in pollen-pistil interaction and pollination biological processes have a lower than expected number of mutations in the coding regions.

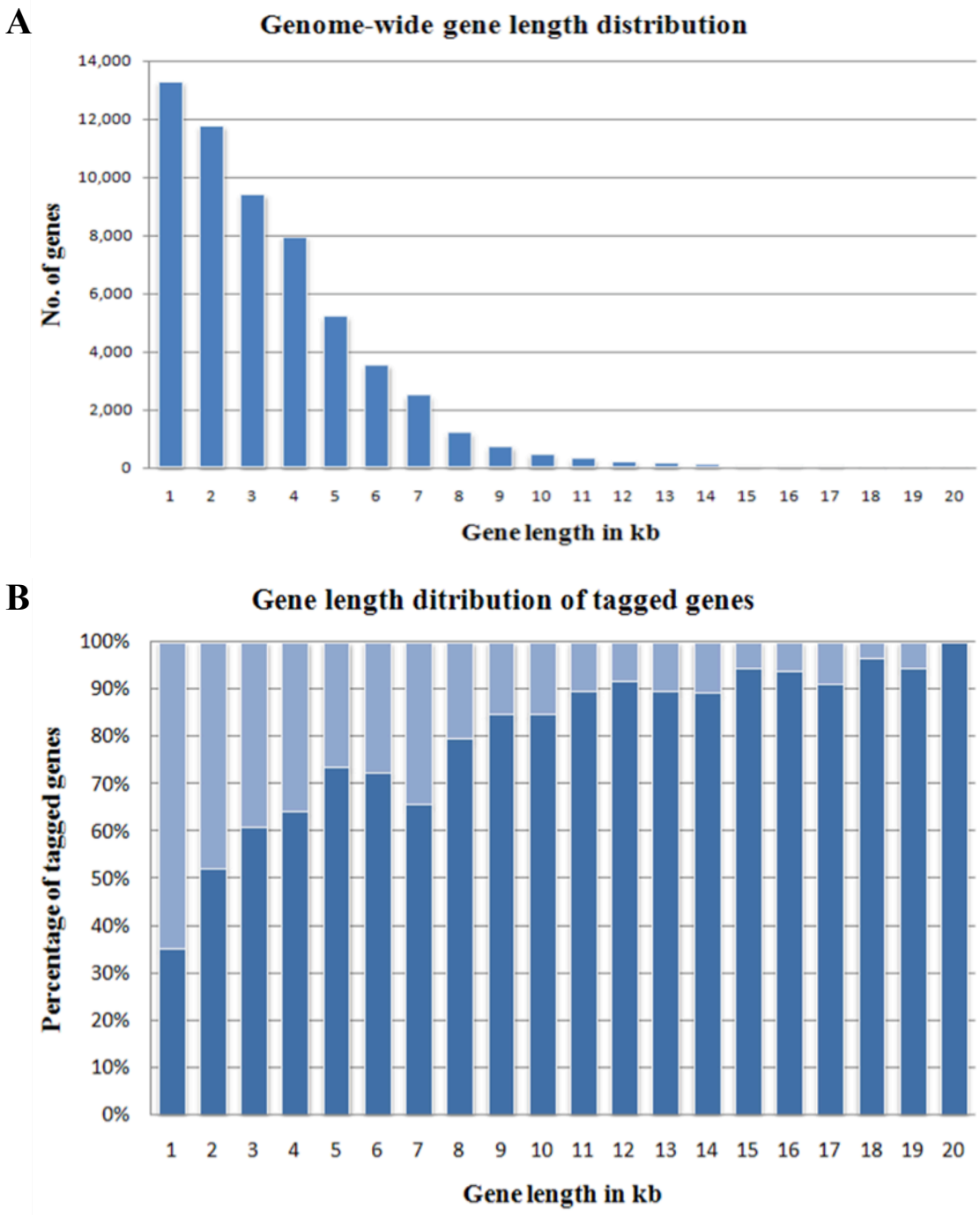


Figure 6.2: Gene size distribution and percentage of genes within each size bin that contain insertion mutations. (**A**) Bar plot of the number of genes in the rice genome that fall within the size bins specified along the x-axis. (**B**) For the same set of size-based bins, the percentages of tagged genes within genes in a particular size bin are plotted with dark blue and the rest of the genes are in light blue.

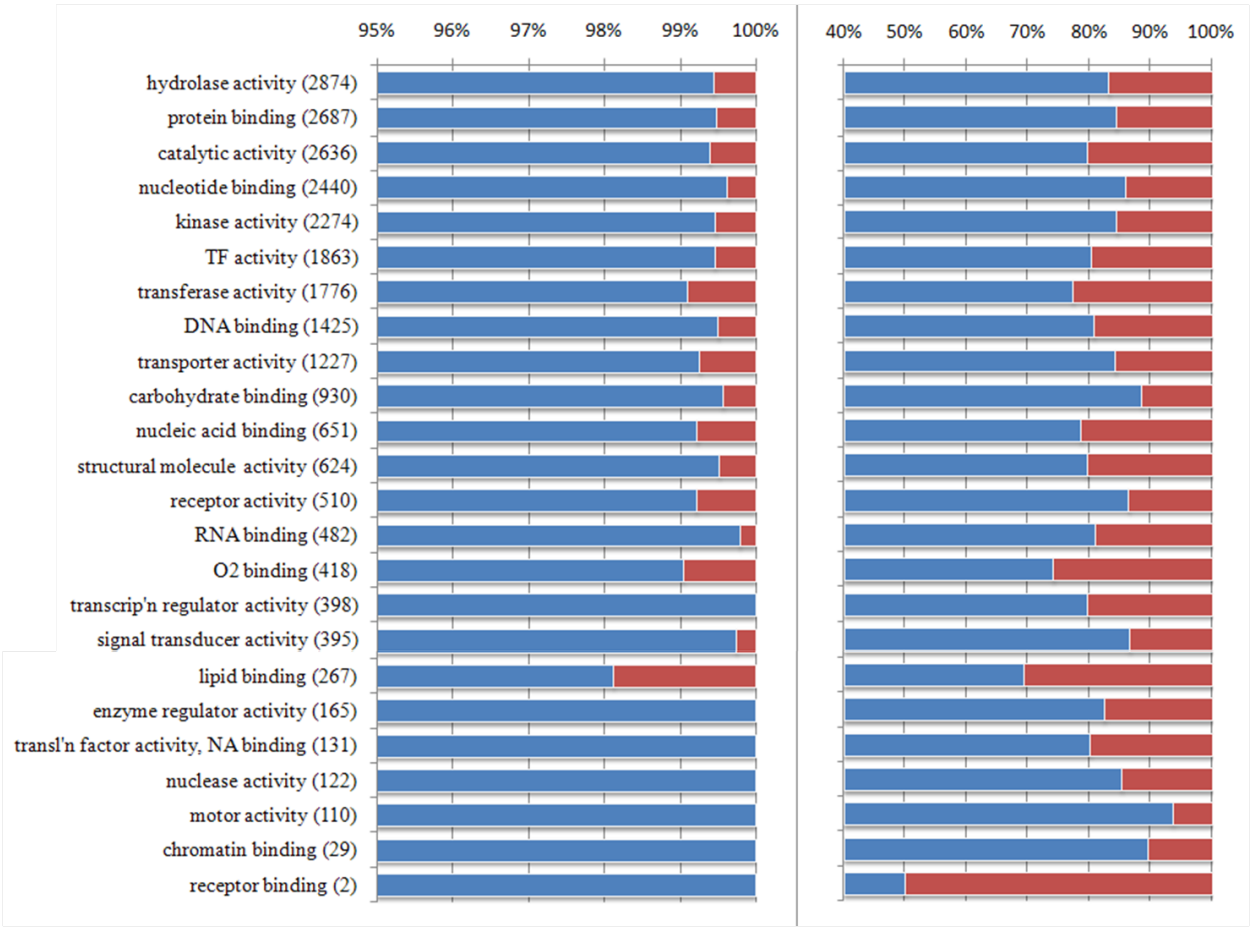


Figure 6.3: GO-slim molecular function categories of genes with inserts (blue), presented as percentage of the total annotated genes (given in brackets) in the category.
The left panel shows the genes with inserts any where in its structure as shown in Figure 1. The right panel indicates the genes with inserts in the exons, introns and 5'UTR as an index of obtaining a knockout mutant. Note that the scales of the x-axes for the right and left panel are different and have been kept this way for clarity. Also to note is that these GO-slim categories are not independent of each other, meaning that different categories could share genes (some time very large, for example, between 'DNA binding' and 'TF activity').

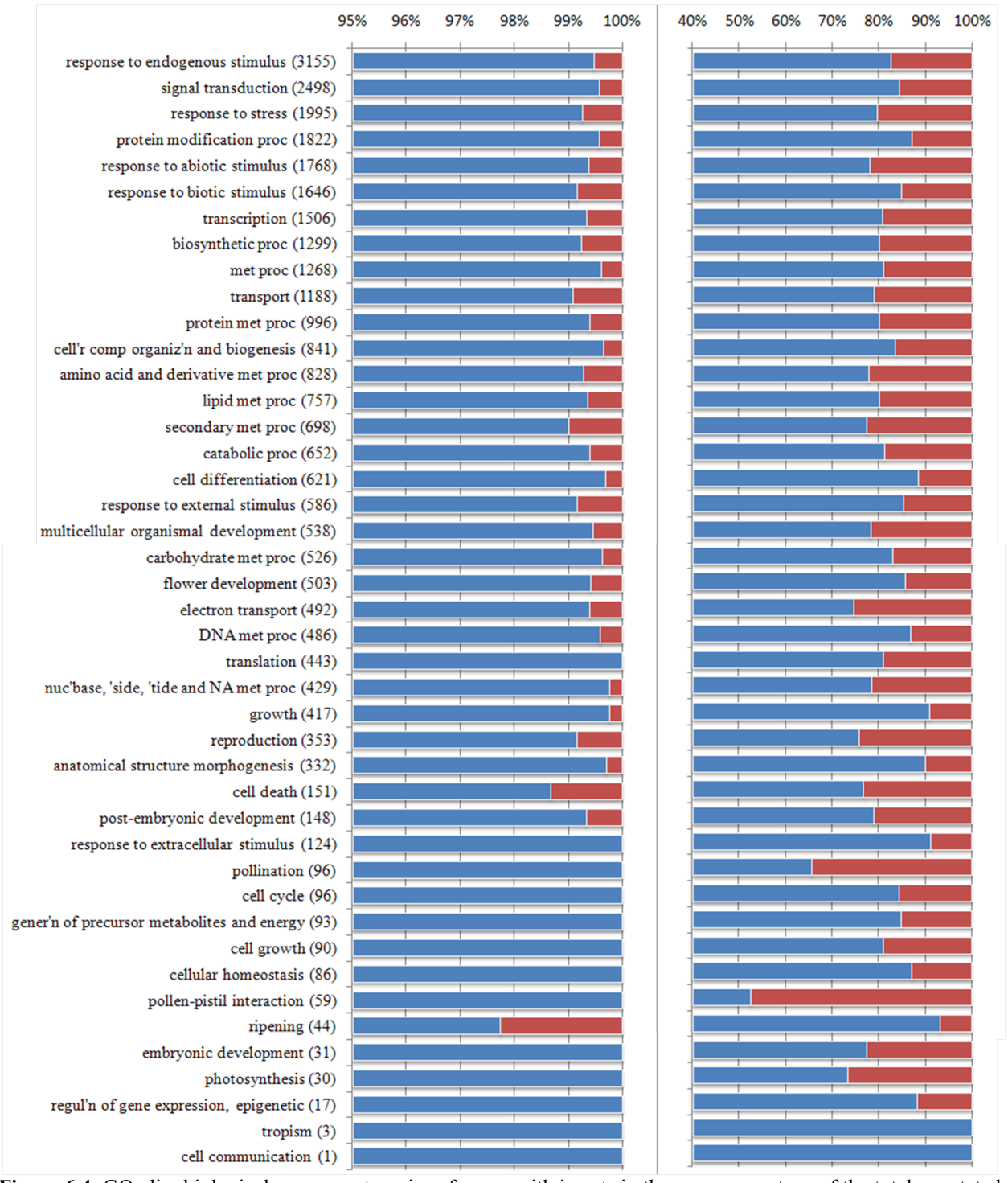


Figure 6.4: GO-slim biological process categories of genes with inserts in them, as percentage of the total annotated genes (given in brackets) in the category.
The left panel shows the genes with inserts, considering the complete gene structure as shown in Figure 1. The right panel indicates the genes with inserts in the exons, introns and 5'UTR as an index of obtaining a knockout mutant. Note that the scales of the x-axes for the right and left panel are different and have been kept this way for clarity. Also to note is that these GO-slim categories are not independent of each other, meaning that different categories could share genes (some time very large, for example, between 'response to stress' and 'response to abiotic stimulus').

6.5. Future development

The size of rice mutant populations generated is large and diverse to suit many functional genomics objectives in the grasses. The number of insertion mutants needed to tag every gene in rice has been estimated to be between 180,698-460,000 (Hirochika et al, 2004). At present, the positions of over 200,000 inserts have been determined by FSTs with KO mutations predicted for 50% of the protein coding genes. Thus, mutants of the remaining genes are required, many of them smaller genes with a lower mutation frequency. Although, the total available mutants are more than two million (Table 6.1), the cataloguing of the mutants by FSTs is limiting, because of the manual manipulations and costs involved. However, new methods of high-throughput sequencing of multi-dimensional DNA pools should be able to assess the genome positions in a more cost-effective way. In addition, the chemical/physical mutagen derived mutants would be accessible by the next generation high-throughput genotyping technologies. For those genes still inaccessible to mutation, probably due to small size, lethality or genome position, more directed gene specific methods using RNAi silencing would be very useful.

6.6. Methods

All the locus ids and coordinates for the whole genome data and the protein-coding subset were downloaded from Gramene BioMart (http://www.gramene.org/biomart/martview/). Of the reported 58,406 total loci and 41,908 protein-coding loci, 56,985 and 41,753 of them, respectively, correspond to nuclear genes with TIGR locus identifiers assigned.

Total no. of reported (genome-wide) loci	**58406**
No. of loci that have TIGR locus ids (LOC_Os[..]g[…..])	**57142**
No. of loci that correspond to nuclear genes	**56985**
No. of loci corresponding to protein-coding genes	**41753**

6.6.1. Datasets

Most of the FST data for inserts from different resources are available from dbGSS (GenBank; http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucgss) and respective project websites. Data for

the present study was obtained from the SIGnAL Rice Functional Genomic Express Database (RiceGE) and through personal communications (see Table 6.S1).

The RiceGE database (http://signal.salk.edu/database/RiceGE) contains recent updates for the following resources, which were used as part of our preliminary dataset of insertion positions for further analysis:

1. AFFRC_NIAS/Tos17
2. CSIRO/Ds
3. OSTID/Ds
4. RIFGP_ZJU/T-DNA
5. RMD/T-DNA
6. SHIP/T-DNA
7. PFG/T-DNA
8. UCD/Ds
9. UCD/dSpm

FST data from CIRAD (T-DNA and *Tos17*), KRDD (*Ds*) and TRIM (T-DNA) were kindly provided by the curators/PIs directly (personal communication).

The Rice genome sequence data and GOslim mappings were obtained from the Rice Genome Annotation Project webpage (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules and http://rice.plantbiology.msu.edu/data_download.shtml respectively). Protein coding sequence assignments were obtained from Gramene (www.gramene.org/).

6.6.2. Analysis

All the FST data was mapped to the Rice genome (v5) using BLAST (Altschul et al, 1990). The results were initially filtered to remove insertion mappings that had an E-value > 1e-10 or those mapping to mitochondrial or chloroplast genomes. Redundancy of mapping position was then removed by considering only one among a group of insertions mapping to the same location in the genome. The final set of insertion positions were used to map onto the chromosome to identify those that correspond to 'genic' and intergenic regions. An insertion was considered to map to a gene if its insertion position fell within the gene itself or within its flanking sequence

(500bp upstream or downstream). The 'genic' insertion positions were further characterized into those that fell within one of the following gene regions: 500bp upstream sequence (promoter), 5' UTR, exon (when not one of the UTRs), intron, 3' UTR, or 500bp downstream sequence.

6.7. References

Agrawal GK, Yamazaki M, Kobayashi M, Hirochika R, Miyao A, Hirochika H (2001) Screening of the rice viviparous mutants generated by endogenous retrotransposon Tos17 insertion. Tagging of a zeaxanthin epoxidase gene and a novel ostatc gene. *Plant Physiol* **125:** 1248-1257

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403-410

An G, Jeong DH, Jung KH, Lee S (2005) Reverse genetic approaches for functional genomics of rice. *Plant Mol Biol* **59:** 111-123

Briggs GC, Osmont KS, Shindo C, Sibout R, Hardtke CS (2006) Unequal genetic redundancies in Arabidopsis--a neglected phenomenon? *Trends Plant Sci* **11:** 492-498

Chen S, Jin W, Wang M, Zhang F, Zhou J, Jia Q, Wu Y, Liu F, Wu P (2003) Distribution and characterization of over 1000 T-DNA tags in rice genome. *Plant J* **36:** 105-113

Chin HG, Choe MS, Lee SH, Park SH, Koo JC, Kim NY, Lee JJ, Oh BG, Yi GH, Kim SC, Choi HC, Cho MJ, Han CD (1999) Molecular analysis of rice plants harboring an Ac/Ds transposable element-mediated gene trapping system. *Plant J* **19:** 615-623

Devos KM (2005) Updating the 'crop circle'. *Curr Opin Plant Biol* **8:** 155-162

Greco R, Ouwerkerk PB, De Kam RJ, Sallaud C, Favalli C, Colombo L, Guiderdoni E, Meijer AH, Hoge Dagger JH, Pereira A (2003) Transpositional behaviour of an Ac/Ds system for reverse genetics in rice. *Theor Appl Genet* **108:** 10-24

Greco R, Ouwerkerk PB, Taal AJ, Sallaud C, Guiderdoni E, Meijer AH, Hoge JH, Pereira A (2004) Transcription and somatic transposition of the maize En/Spm transposon system in rice. *Mol Genet Genomics* **270:** 514-523

Hirochika H, Guiderdoni E, An G, Hsing YI, Eun MY, Han CD, Upadhyaya N, Ramachandran S, Zhang Q, Pereira A, Sundaresan V, Leung H (2004) Rice mutant resources for gene discovery. *Plant Mol Biol* **54:** 325-334

Hsing YI, Chern CG, Fan MJ, Lu PC, Chen KT, Lo SF, Sun PK, Ho SL, Lee KW, Wang YC, Huang WL, Ko SS, Chen S, Chen JL, Chung CI, Lin YC, Hour AL, Wang YW, Chang YC, Tsai MW et al (2007) A rice gene activation/knockout mutant resource for high throughput functional genomics. *Plant Mol Biol* **63:** 351-364

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436:** 793-800

Jeon JS, Lee S, Jung KH, Jun SH, Jeong DH, Lee J, Kim C, Jang S, Yang K, Nam J, An K, Han MJ, Sung RJ, Choi HS, Yu JH, Choi JH, Cho SY, Cha SS, Kim SI, An G (2000) T-DNA insertional mutagenesis for functional genomics in rice. *Plant J* **22:** 561-570

Jeong DH, An S, Kang HG, Moon S, Han JJ, Park S, Lee HS, An K, An G (2002) T-DNA insertional mutagenesis for activation tagging in rice. *Plant Physiol* **130:** 1636-1644

Jeong DH, An S, Park S, Kang HG, Park GG, Kim SR, Sim J, Kim YO, Kim MK, Kim J, Shin M, Jung M, An G (2006) Generation of a flanking sequence-tag database for activation-tagging lines in japonica rice. *Plant J* **45:** 123-132

Jung KH, Hur J, Ryu CH, Choi Y, Chung YY, Miyao A, Hirochika H, An G (2003) Characterization of a rice chlorophyll-deficient mutant using the T-DNA gene-trap system. *Plant Cell Physiol* **44:** 463-472

Kang HG, Park S, Matsuoka M, An G (2005) White-core endosperm floury endosperm-4 in rice is generated by knockout mutations in the C-type pyruvate orthophosphate dikinase gene (OsPPDKB). *Plant J* **42:** 901-911

Kolesnik T, Szeverenyi I, Bachmann D, Kumar CS, Jiang S, Ramamoorthy R, Cai M, Ma ZG, Sundaresan V, Ramachandran S (2004) Establishing an efficient Ac/Ds tagging system in rice: large-scale analysis of Ds flanking sequences. *Plant J* **37:** 301-314

Kumar CS, Wing RA, Sundaresan V (2005) Efficient insertional mutagenesis in rice using the maize En/Spm elements. *Plant J* **44:** 879-892

Lee S, Kim J, Son JS, Nam J, Jeong DH, Lee K, Jang S, Yoo J, Lee J, Lee DY, Kang HG, An G (2003) Systematic reverse genetic screening of T-DNA tagged genes in rice for functional genomic analyses: MADS-box genes as a test case. *Plant Cell Physiol* **44:** 1403-1411

Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, Abe K, Shinozuka Y, Onosato K, Hirochika H (2003) Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell* **15:** 1771-1780

Parinov S, Sundaresan V (2000) Functional genomics in Arabidopsis: large-scale insertional mutagenesis complements the genome sequencing project. *Curr Opin Biotechnol* **11:** 157-161

Park SH, Jun NS, Kim CM, Oh TY, Huang J, Xuan YH, Park SJ, Je BI, Piao HL, Cha YS, Ahn BO, Ji HS, Lee MC, Suh SC, Nam MH, Eun MY, Yi G, Yun DW, Han CD (2007) Analysis of gene-trap Ds rice populations in Korea. *Plant Mol Biol* **65:** 373-384

Pereira A (2000) A transgenic perspective on plant functional genomics. *Transgenic Res* **9:** 245-260; discussion 243

Piffanelli P, Droc G, Mieulet D, Lanau N, Bes M, Bourgeois E, Rouviere C, Gavory F, Cruaud C, Ghesquiere A, Guiderdoni E (2007) Large-scale characterization of Tos17 insertion sites in a rice T-DNA mutant library. *Plant Mol Biol* **65:** 587-601

Qu S, Desai A, Wing R, Sundaresan V (2008) A versatile transposon-based activation tag vector system for functional genomics in cereals and other monocot plants. *Plant Physiol* **146:** 189-199

Sallaud C, Gay C, Larmande P, Bes M, Piffanelli P, Piegu B, Droc G, Regad F, Bourgeois E, Meynard D, Perin C, Sabau X, Ghesquiere A, Glaszmann JC, Delseny M, Guiderdoni E (2004) High throughput T-DNA insertion mutagenesis in rice: a first step towards in silico reverse genetics. *Plant J* **39:** 450-464

Sallaud C, Meynard D, van Boxtel J, Gay C, Bes M, Brizard JP, Larmande P, Ortega D, Raynal M, Portefaix M, Ouwerkerk PB, Rueb S, Delseny M, Guiderdoni E (2003) Highly efficient production and characterization of T-DNA plants for rice (Oryza sativa L.) functional genomics. *Theor Appl Genet* **106:** 1396-1408

Sterck L, Rombauts S, Vandepoele K, Rouze P, Van de Peer Y (2007) How many genes are there in plants (... and why are they there)? *Curr Opin Plant Biol* **10:** 199-203

Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, Ma H, Martienssen R (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev* **9:** 1797-1810

Takano M, Kanegae H, Shinomura T, Miyao A, Hirochika H, Furuya M (2001) Isolation and characterization of rice phytochrome A mutants. *Plant Cell* **13:** 521-534

Till BJ, Cooper J, Tai TH, Colowit P, Greene EA, Henikoff S, Comai L (2007) Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biol* **7:** 19

Upadhyaya NM, Zhou XR, Zhu QH, Ramm K, Wu LM, Eamens A, Sivakumar R, Kato T, Yun DW, Santhoshkumar C, Narayanan KK, Peacock JW, Dennis ES (2002) An iAc/Ds gene and enhancer trapping system for insertional mutagenesis in rice. *Functional Plant Biology* **29:** 547-559

Upadhyaya NM, Zhu QH, Zhou XR, Eamens AL, Hoque MS, Ramm K, Shivakkumar R, Smith KF, Pan ST, Li S, Peng K, Kim SJ, Dennis ES (2006) Dissociation (Ds) constructs, mapped Ds launch pads and a transiently-expressed transposase system suitable for localized insertional mutagenesis in rice. *Theor Appl Genet* **112:** 1326-1341

van Enckevort LJ, Droc G, Piffanelli P, Greco R, Gagneur C, Weber C, Gonzalez VM, Cabot P, Fornara F, Berri S, Miro B, Lan P, Rafel M, Capell T, Puigdomenech P, Ouwerkerk PB, Meijer AH, Pe E, Colombo L, Christou P et al (2005) EU-OSTID: a collection of transposon insertional mutants for functional genomics in rice. *Plant Mol Biol* **59:** 99-110

Wu JL, Wu C, Lei C, Baraoidan M, Bordeos A, Madamba MR, Ramos-Pamplona M, Mauleon R, Portugal A, Ulat VJ, Bruskiewich R, Wang G, Leach J, Khush G, Leung H (2005) Chemical- and irradiation-induced mutants of indica rice IR64 for forward and reverse genetics. *Plant Mol Biol* **59:** 85-97

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, Zhang J, Zhang Y, Li R, Xu Z, Li X, Zheng H, Cong L, Lin L, Yin J, Geng J et al (2005) The Genomes of Oryza sativa: a history of duplications. *PLoS Biol* **3:** e38

Zhang J, Li C, Wu C, Xiong L, Chen G, Zhang Q, Wang S (2006) RMD: a rice mutant database for functional analysis of the rice genome. *Nucleic Acids Res* **34:** D745-748

Zhu QH, Eun MY, Han CD, Kumar CS, Pereira A, Ramachandran S, Sundaresan V, Eamens AL, Upadhyaya NM, Wu R (2007) Transposon insertional mutants: a resource for rice functional genomics. In *Rice Functional Genomics—Challenges, Progress and Prospects*, Upadhyaya NM (ed), pp 223–271. New York: Springer

Zhu QH, Hoque MS, Dennis ES, Upadhyaya NM (2003) Ds tagging of BRANCHED FLORETLESS 1 (BFL1) that mediates the transition from spikelet to floret meristem in rice (Oryza sativa L). *BMC Plant Biol* **3:** 6

Zhu QH, Ramm K, Shivakkumar R, Dennis ES, Upadhyaya NM (2004) The ANTHER INDEHISCENCE1 gene encoding a single MYB domain protein is involved in anther development in rice. *Plant Physiol* **135:** 1514-1525

7. A resource for systems analysis of transcriptional modules involved in drought response in rice

Arjun Krishnan, Madana Ambavaram, Utlwang Batlang, Andy Pereira

7.1. Abstract

Water scarcity for crop production can impose abiotic stresses such as drought and salinity, which together with other environmental stress factors can cause depreciation in crop yield up to 70%. The severity of impact of drought on crops like rice is contingent on the developmental stage of the plant with the most sensitive stage being the reproductive phase. Hence, functional genomic approaches, mainly gene expression profiling, have been used to understanding drought response and resistance mechanisms in rice. Here we present the drought transcriptomes of rice in three developmental stages and gain insights into the processes and regulatory mechanisms involved in common and stage-specific drought responses. Moreover, we have created a resource for systems analysis of drought response in rice that allows exploration of transcriptional modules, and associated regulatory, functional and genomic annotations. Most importantly, it aids in choosing nonobvious candidates for genes involved in mediating drought response and hence in drought tolerance. This flexible approach can be extended to all the other abiotic stresses with ease and should aid in engineering drought tolerance.

7.2. Introduction

Drought is one of the most widespread of environmental stresses that has a large impact on crop growth and yield. Around 70% of our total fresh water resources is used in agriculture, and rice, a major food crop, consumes 30% of all the fresh water used in agriculture. Due to this dependence, water scarcity for crop production can impose drought that, together with other environmental stress factors, can cause depreciation in crop yield up to 70% (Boyer, 1982). Plants have intrinsically developed drought resistance mechanisms for protection at various stages of development and under varying environmental conditions. Plants confronted by drought thus exhibit drought responses that can lead to the induction of diverse protective mechanisms of resistance at every organ and cellular levels (Dinneny et al, 2008). Hence, there

are various developmental windows and stress responses where specific plant organs are protected against dehydration and drought that may follow some general mechanisms of coordinate gene action. It follows that these general stress mechanism modules have been acquired in the evolution of plants to act in specific situations and environments. This hypothesis is supported by studies where genes normally expressed in specific developmental stages (Aharoni et al, 2004) or stress induced conditions (Kasuga et al, 1999) can confer drought resistance when overexpressed in plants.

While most studies have considered drought responses and resistance mechanisms at the vegetative growth stage, the major cereals like rice and maize are most sensitive at the reproductive stages involved in pollen fertility and grain development. In rice, anther dehiscence and panicle exsertion are well known events sensitive to drought, respectively showing irreversible and partly reversible disruption upon stress (Liu et al, 2006). Water deficit shortly after fertilization can cause ovary abortion due to low sugar signaling that down-regulates sucrose processing invertases. Invertase is essential in sink tissue, such as the maize ovary, undergoing cell division and growth (Andersen et al, 2002), and also for peduncle development and anthesis in rice (Ji et al, 2005a). This regulation of invertase activity and transcription constitute known mechanistic bases for drought sensitivity in cereal reproduction systems.

The response of plants to drought stress is extremely complex and spans several orders of magnitude in time and space, causing system-wide protective responses and adverse reactions. Gene expression profiling has been used often to capture a facet of the cellular response to drought and other abiotic stresses (Deyholos, 2010). Due to the accumulation of large-scale gene expression datasets measuring gene expression across diverse conditions in plants, building and understanding the transcriptional organization of the cell using coexpression networks has become a wide-spread and powerful approach (Usadel et al, 2009). Particularly, in rice, recent studies have used gene coexpression to gain biological insights into general (Wang et al, 2009) and case-specific (Fu & Xue, 2010) gene regulation. In addition, other studies have brought together coexpression network resources for rice functional genomics (Lee et al, 2009; Ogata et al, 2010). Based on the tenet that cellular organization is modular (Hartwell et al, 1999), identification of dense clusters or modules in coexpression networks has been used as an

approach to discover functional modules that are groups of genes/proteins that work together to perform a coherent biological function inside the cell (Ficklin et al, 2010; Fukushima et al, 2009).

Typically, for coexpression analysis, all the expression data across several diverse conditions are amalgamated into one large matrix that is used for calculation of expression correlation between genes across all the conditions. However, it is has observed that pairs of genes could have varying expression dynamics in different conditions and hence, calculating conditional coexpression is crucial in understanding specific correlations (Kostka & Spang, 2004; Rawat et al, 2008). None of the coexpression studies in rice have explored conditional coexpression and analyzed the network at a modular level on a genome-scale. In this work, we present a resource for comprehensive analysis of drought stress in rice on the basis of an underlying modular coexpression network specific to environment condition response. First, we performed gene expression profiling of rice plants subjected to drought at three growth stages, followed by functional analysis and regulatory sequence discovery. Then, to take the analysis to a network-level, we integrated publicly available rice gene expression datasets generated in the context of response to some environmental condition. We constructed, what is termed, the Rice Environment Coexpression Network or RECoN, based on gene expression correlation, and partitioned RECoN into dense clusters. Finally, from all the clusters, we teased out drought-related clusters using drought-response genes identified from our experiments. By interfacing these clusters with functional annotation, regulatory sequence information, quantitative trait loci (QTLs) and mutant resources in rice, we thus created a large framework for exhaustive exploration of drought response and tolerance in rice.

7.3. Results

7.3.1. Gene expression profiling of progressive drought in rice

In order to gain a global picture of drought stress and its effects on the plant at very specific drought sensitive growth stages, we profiled samples from rice plants to a progressive drought treatment at the seedling, vegetative and reproductive stages using the rice Affymetrix GeneChips. Based on statistical analysis of differential expression, we observed that a large number of genes are perturbed given a stringent q-value cut-off of <0.01 was used (Table 7.S1).

The largest shift in expression compared to well-watered controls happened at the seedling stage with ~12,300 genes showing differential expression, compared to only ~2,500 genes in the reproductive stage. Drought response at the vegetative stage involved changes in ~9000 genes. On comparing the number of genes shared between the response at the three stages, we found that on an average ~55% of each set of genes was shared with the genes in the other stages in the case of both up- and down-regulated genes (Fig. 7.1).

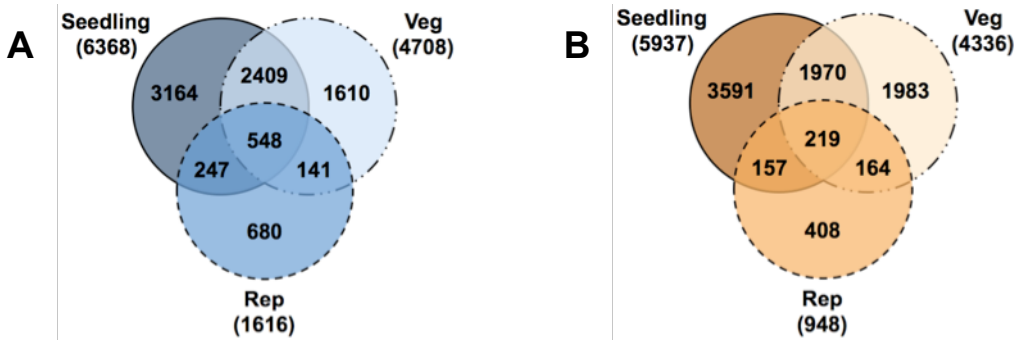
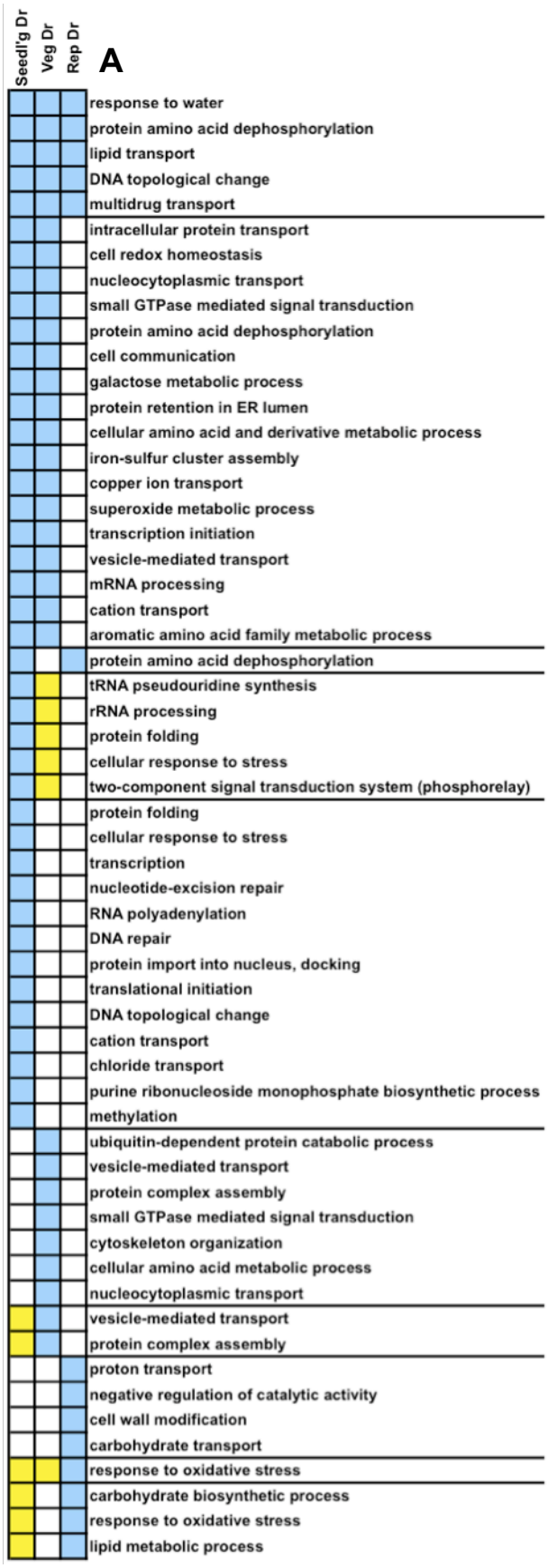


Figure 7.1: Gene expression profiles under drought.
Venn diagrams comparing up-regulated (**A**) and down-regulated (**B**) genes in response to drought in three growth stages: seedling, vegetative and reproductive. Total numbers of genes for all gene sets are indicated in brackets.

7.3.2. Biological processes and regulatory elements involved in drought-regulated gene expression in rice

The foremost aspects of drought response here are the common and the stage-specific responses. More than comparing the number of genes, we first took a union of all the drought-regulated genes and split them into sets of genes that show identical pattern of regulation across the stages. Then, we determined processes defined by Gene Ontology (GO) biological process (BP) annotations that were enriched in each of these gene sets (Table 7.S2). As expected, the most significant GO term among the set of genes up-regulated in all stages was 'response to water'. Similarly, different combinations of genes involved in protein dephosphorylation and small GTPase-mediated signaling (Mazzucotelli et al, 2008) are up-regulated in all stages. Among the genes down-regulated in all stages, those involved in photosynthesis and related processes were enriched. Genes involved in translation were induced and repressed in the seedling and reproductive stages, respectively. On the other hand, genes involved in cell wall modification, which are known to be differently regulated depending on the species, organ, and tissue (Moore et al, 2008), are repressed at the seedling stage but up-regulated at the reproductive stage.


Seedl'g Dr
Veg Dr
Rep Dr
A
response to water
protein amino acid dephosphorylation
lipid transport
DNA topological change
multidrug transport
intracellular protein transport
cell redox homeostasis
nucleocytoplasmic transport
small GTPase mediated signal transduction
protein amino acid dephosphorylation
cell communication
galactose metabolic process
protein retention in ER lumen
cellular amino acid and derivative metabolic process
iron-sulfur cluster assembly
copper ion transport
superoxide metabolic process
transcription initiation
vesicle-mediated transport
mRNA processing
cation transport
aromatic amino acid family metabolic process
protein amino acid dephosphorylation
tRNA pseudouridine synthesis
rRNA processing
protein folding
cellular response to stress
two-component signal transduction system (phosphorelay)
protein folding
cellular response to stress
transcription
nucleotide-excision repair
RNA polyadenylation
DNA repair
protein import into nucleus, docking
translational initiation
DNA topological change
cation transport
chloride transport
purine ribonucleoside monophosphate biosynthetic process
methylation
ubiquitin-dependent protein catabolic process
vesicle-mediated transport
protein complex assembly
small GTPase mediated signal transduction
cytoskeleton organization
cellular amino acid metabolic process
nucleocytoplasmic transport
vesicle-mediated transport
protein complex assembly
proton transport
negative regulation of catalytic activity
cell wall modification
carbohydrate transport
response to oxidative stress
carbohydrate biosynthetic process
response to oxidative stress
lipid metabolic process

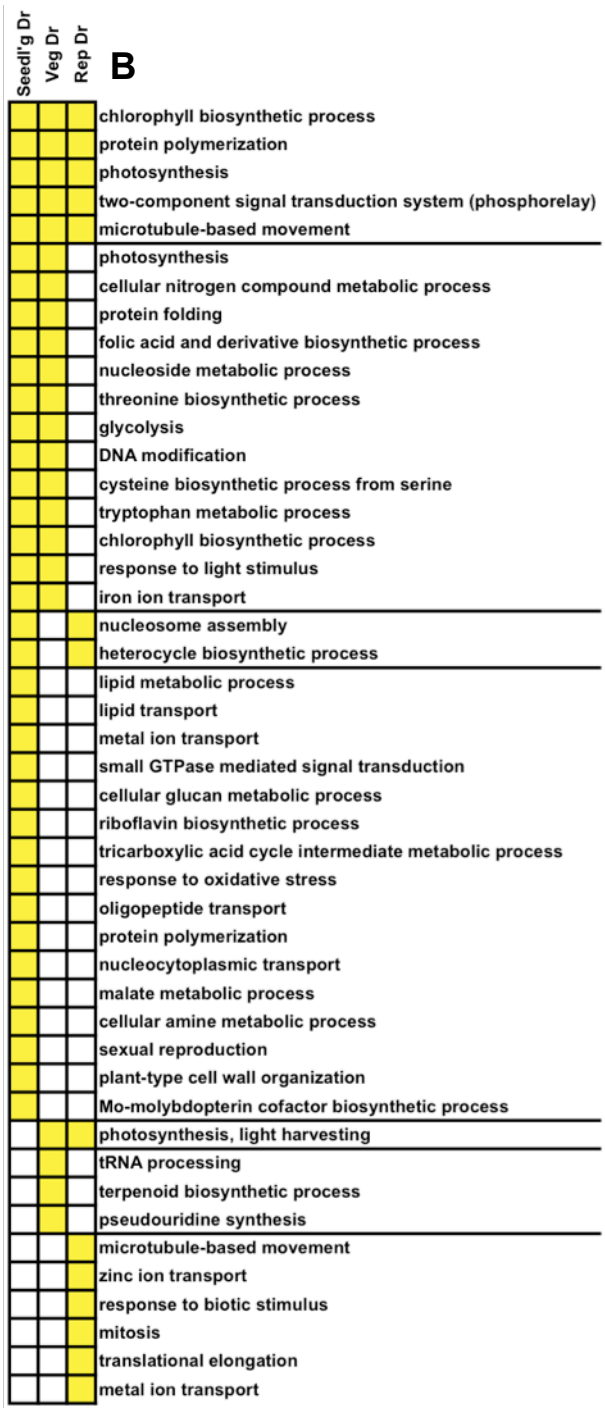

Seedl'g Dr
Veg Dr
Rep Dr
B
chlorophyll biosynthetic process
protein polymerization
photosynthesis
two-component signal transduction system (phosphorelay)
microtubule-based movement
photosynthesis
cellular nitrogen compound metabolic process
protein folding
folic acid and derivative biosynthetic process
nucleoside metabolic process
threonine biosynthetic process
glycolysis
DNA modification
cysteine biosynthetic process from serine
tryptophan metabolic process
chlorophyll biosynthetic process
response to light stimulus
iron ion transport
nucleosome assembly
heterocycle biosynthetic process
lipid metabolic process
lipid transport
metal ion transport
small GTPase mediated signal transduction
cellular glucan metabolic process
riboflavin biosynthetic process
tricarboxylic acid cycle intermediate metabolic process
response to oxidative stress
oligopeptide transport
protein polymerization
nucleocytoplasmic transport
malate metabolic process
cellular amine metabolic process
sexual reproduction
plant-type cell wall organization
Mo-molybdopterin cofactor biosynthetic process
photosynthesis, light harvesting
tRNA processing
terpenoid biosynthetic process
pseudouridine synthesis
microtubule-based movement
zinc ion transport
response to biotic stimulus
mitosis
translational elongation
metal ion transport

Figure 7.2: Functions and processes common and specific to various drought stress treatments and time-points. These are defined broadly based on Gene Ontology (GO) biological process (BP) annotations of rice genes. First, the total of all drought-regulated genes from all stages were pooled together and were then partitioned based on the combination of their regulation in the three stages (e.g. up-up-up, or down-up-down). Then, GO BP terms of interest (rows) were identified by analysis of enrichment of the set of genes annotated with a given GO BP term in each regulation-combination defined by the yellow-blue color-coding along the rows where blue means up-regulation and yellow means down-regulation. Statistical significance of enrichment was calculated using the hypergeometric test and terms with *q-value* <0.1 in at least one of the treatments were retained. (**A**) GO BP terms enriched in gene sets up-regulated in at least one stage. (**B**) GO BP terms enriched in gene sets only down-regulated in one or more stage.

Although this analysis gave us a few insights into drought-regulated gene expression, apart from the previously known stress response themes, it is hard to pinpoint biological functions that get affected specifically in the different stages. The most important reason for this shortfall is the fact that rice genes are extremely poorly characterized and very few genes have been annotated well in public databases while several insights about individual genes can be teased out by plodding through the literature. This is scenario becomes evident when we look at the small number of genes common between any GO term and the set of drought genes (Table 7.S2). Therefore, we need other approaches to pursue that will give us a better picture of the underlying changes during drought.

The gene expression changes observed in such profiling experiments could be due to transcriptional or post-transcriptional regulation (especially by small RNAs). In order to decipher the transcriptional regulation underlying the differential expression, we turned to analysis of cis-regulatory elements (CREs). Based on the two pieces of good information we have – gene differential expression and genome sequence data – we devised a *de novo* CRE discovery pipeline. We chose a *de novo* approach because our knowledge about verified regulatory sequences is again poor in rice and there is enormous room for identification of novel CREs that mediate gene regulation under various conditions including drought. Using the *de novo* motif discovery tool FIRE (Elemento et al, 2007) we discerned short DNA sequences in the up-stream regions of genes that are informative about the expression pattern of the genes (see *Methods*). We subsequently compared the newly identified motifs to known cis-elements in the PLACE database of known cis-regulatory elements (Higo et al, 1999) using STAMP (Mahony & Benos, 2007). Applying this procedure to drought-regulated genes in the three stages (further separated into up- and down-regulated gene sets) led to the identification of known and novel CREs. The discovered motifs, the drought-response gene sets they were discovered in and comparison to

known motifs are presented in Figure 7.3. Motifs discussed below are referred to by the name of the most similar 'known' CRE (see the 'PLACE motifs' table in Figure 3 for the key).

Motifs similar to the ABRE (Hattori et al, 2002) were identified among genes up-regulated in all the stages and could, hence, mediate the up-regulation of the stress response genes. On the other hand, two different motifs – SORLIP (Jiao et al, 2005) and Ibox (Giuliano et al, 1988) – were found in genes down-regulated at the seedling and vegetative stages, respectively, that have been associated with light-induced gene expression (e.g. related to photosynthesis). Although photosynthesis-related genes are also repressed in the reproductive stage, no known light-associated motif was found in the repressed genes except for a novel motif HAGCTAVCD. Based on this result, we propose that while the induction of typical water deficit response genes is similarly mediated by the ABRE-like CREs in all stages, there is a substantial difference in the repression of the photosynthesis-related genes mediated by three different CRE. Intriguingly, another element involved in light-mediated gene activation – Tbox (Chan et al, 2001) – was found among genes up-regulated by drought at the seedling stage.

A Telo-box-like motif was identified in the genes up-regulated at the seedling stage. The Telo-box has been confirmed to be important in the regulation of ribosomal protein genes (Tremousaygue et al, 2003) and there is some evidence that these genes are up-regulated by drought at the seedling stage. MYB transcription factor (TF) binding sites were found among upstream sequences of genes down-regulated at the seedling stage and up-regulated at the vegetative stage. Moreover, four distinct novel motifs have been discovered associated with drought at all three stages.

De novo Motif ID	Seedl'g Dr	Veg Dr	Rep Dr	De novo Motif Sequence	Z-score	PLACE motifs					
						ID	Sequence	E-value	ID	Sequence	E-value
Seedling_Dr_m1				RAACCCTM	163	UP2ATMSD	AAACCCTA	5.77E-11	TELOBOXATEEF1AA1	AAACCCTAA	5.10E-10
Seedling_Dr_m4				CAAAGCC	13.4	TBOXATGAPB	ACTTTG	6.24E-06			
Seedling_Dr_m5				WAAAAAAAS	56.1	minus314MOTIFZMSBE1	ACATAAAATAAAAAAAGGCA	4.43E-09	MARTBOX	TTWTWTTWTT	3.05E-08
Seedling_Dr_m8				VACCAAMC	49.6	MYBPLANT	MACCWAMC	5.77E-11	ACIIPVPAL2	CCACCAACCCCC	9.19E-08
Veg_Dr_m2				MACSAMAC	64	MYBPLANT	MACCWAMC	1.08E-07			
Seedling_Dr_m6				DYTAGCTAV	83	SE2PVGRP1	TTNNNGTAGCTAGTGTATTTGTAT	5.56E-06			
Rep_Dr_m3				HAGCTAVCD	36.9						
Veg_Dr_m5				YCYTATCC	55.6	IBOXCORENT	GATAAGR	8.77E-10			
Seedling_Dr_m2				YHCGTSTC	109	ABREMOTIFAOSOSEM	TACGTGTC	2.50E-08			
Rep_Dr_m1				BRCGTGTC	36.9	ABREMOTIFAOSOSEM	TACGTGTC	5.96E-10			
Veg_Dr_m1				CKCCTCV	245.7	ABRECE3ZMRAB28	ACGCGCCTCCTC	7.63E-06			
Seedling_Dr_m7				CWCACTSY	55.5	SORLIP5AT	GAGTGAG	3.98E-08			
Veg_Dr_m4				ANTCGCA	17						
Veg_Dr_m3				HTCYGGAAR	25.3						
Seedling_Dr_m3				BTCTSGAA	48.9						
Rep_Dr_m2				TCCAGWMV	20.7						

Figure 7.3: Cis-regulatory elements identified in the upstream regions of drought-regulated genes from the three growth stages.
Each element identified along the rows was identified using *de novo* motif discovery to identify short degenerate DNA sequences whose presence or absence in the 1Kb upstream regions of genes is highly informative about the expression of the given gene set (e.g. up-regulated genes in Reproductive stage) given the background distribution of the sequence in the upstream sequences of all the genes in the genome. The colored matrix indicates which motifs were identified using genes regulated in which stage, again with yellow indicating down-regulation and blue up-regulation. Motifs informative about up- and down-regulation together are indicated by green. In the adjoining table, the sequence of the *de novo* motifs are given in the nucleotide IUPAC nomenclature along with the Z-score of the information value of the motif reflecting how far the observed value is, in number of standard deviations, from the average random information (see *Methods*). These motifs were then compared to known cis-elements in the PLACE database using the STAMP web server. Known elements with significant match to each *de novo* motif are presented in the 'PLACE motifs' table in the form of the database ID, DNA sequence and E-value of sequence match with the *de novo* motif. Motifs with no match to any known element are novel putative regulatory elements.

7.3.3. Drought-response clusters in the environment coexpression network of rice

Next, we wished to further dissect the genes identified using differential analysis by splitting them into groups of genes that might work together to perform a common function or carry out a specific biological process. Coexpression networks have been used extensively in plants to organize genes into expression-based modules and explore their functions (Mao et al, 2009; Mentzen & Wurtele, 2008). In addition, since the CRE analysis identified motifs that differentiate groups of genes with different expression patterns, it is more intuitively correct to break down the large sets of differentially expressed genes into finer-scale coherent modules of genes on which one could redo the analysis. This is also true in the case of functional enrichment analysis. Furthermore, comparison between responses at different stages might be more meaningful and robust at the modular level than at the level of individual genes due to several factors including different subsets of the same cellular apparatus being used/regulated in different stages, and noisy high-throughput analysis increasing both false-positive and false-negative gene-gene overlaps. Therefore, we use publicly available gene expression data to aid in determining biologically meaningful transcriptional modules in rice, which could then be used to delineate modules relevant to drought (based on the drought experiments presented in this study).

To do this, we designed an extensive pipeline that uses data from our drought experiments and publicly available gene expression profiles to elucidate clusters of densely connected genes relevant to drought (Fig. 7.4.) and interface it with regulatory, functional and genomic annotations. Drought-responsive genes were carried over from previous analysis (Fig. 7.4, step 1a). In parallel, we obtained 129 publicly available rice Affymetrix microarrays related to

response of the rice plant to some environmental condition and worked with the raw data (Fig. 7.4, step 1b). We normalized and summarized the data into a gene expression matrix based on a custom probe-gene reannotation of the rice GeneChip. The reannotation increases the accuracy of the gene expression quantification process by assigning only specific probes to genes, and increases coverage of the array. We then converted the gene expression data into a matrix of 34,792 genes and 45 distinct conditions/groups and used it to construct a coexpression network connecting pairs of genes that have a significantly high correlation between their expression profiles across the conditions (top 2.5% of all pairs of genes ordered in decreasing order of correlation; see *Methods*). This network, termed Rice Environment Coexpression Network or RECoN, contains 34,792 genes connected by ~18.5 million edges.

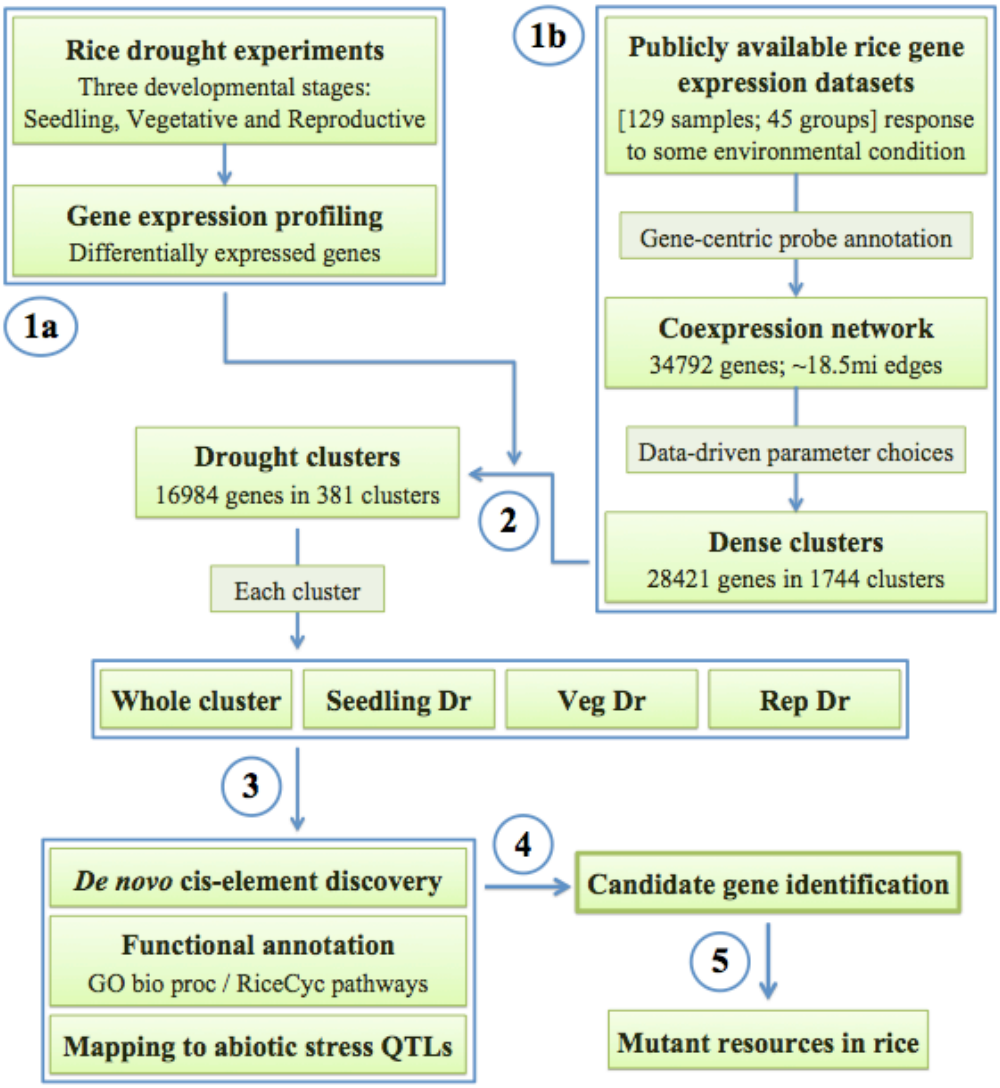


Figure 7.4: Workflow for mining and characterization of drought transcriptional modules.
(**1a**) Identification of drought-responsive genes in the three growth stages. (**1b**) Reconstruction and clustering of the rice environmental coexpression network from publicly available gene expression datasets. (**2**) Determination of 'drought' clusters based on the combination of results from the steps 1a and 1b, and extraction of whole cluster and specific drought gene sets (**3**) Functional enrichment analysis, cis-regulatory motif discovery and mapping to known abiotic stress QTL intervals in the rice genome. (**4**) Presentation of these data to the user where (s)he explores the results to identify gene candidates for experimental verification. (**5**) Information on availability of mutants in genes of interest that can be used to study gene function.

There are several clustering algorithms that work with weighted networks and find groups of densely connected nodes (Bader & Hogue, 2003; Enright et al, 2002). SPICi is a recently developed clustering tool whose strength lies in its ability to cluster large networks extremely fast, still maintaining a level of performance comparable to previous state-of-the-art algorithms (Jiang & Singh, 2010). Therefore, we used SPICi to cluster our extremely large network. However, like every clustering algorithm, amongst a few, there is a single user defined parameter T_d that determines the density of the resultant clusters and heavily influences the clustering process. To avoid an ad hoc or even a wrong choice of this parameter, we performed exhaustive data-driven tests on the network clustered using a range of T_d values to identify the best parameter for the network at hand (Fig. 7.5).

First, for different values of T_d, we tracked the number of clusters obtained and the fraction of genes in the original network that were in clusters of 3 or more genes (Fig. 7.5A). At small values of T_d there are very few clusters and only a few broken links. As T_d increases, the number of clusters increases, but, however, very high T_d will break the network so much that the clusters with 3 or more genes will again become rare. Similarly, as T_d increases, the number of genes that are part of clusters will steady decrease until a critical value beyond which a large portion of genes will get disconnected and fall out of good-sized clusters. Therefore, looking for the value of T_d after which there is the first significant drop in the number of clusters and fraction of genes in clusters, we found that this happens at T_d=0.65.

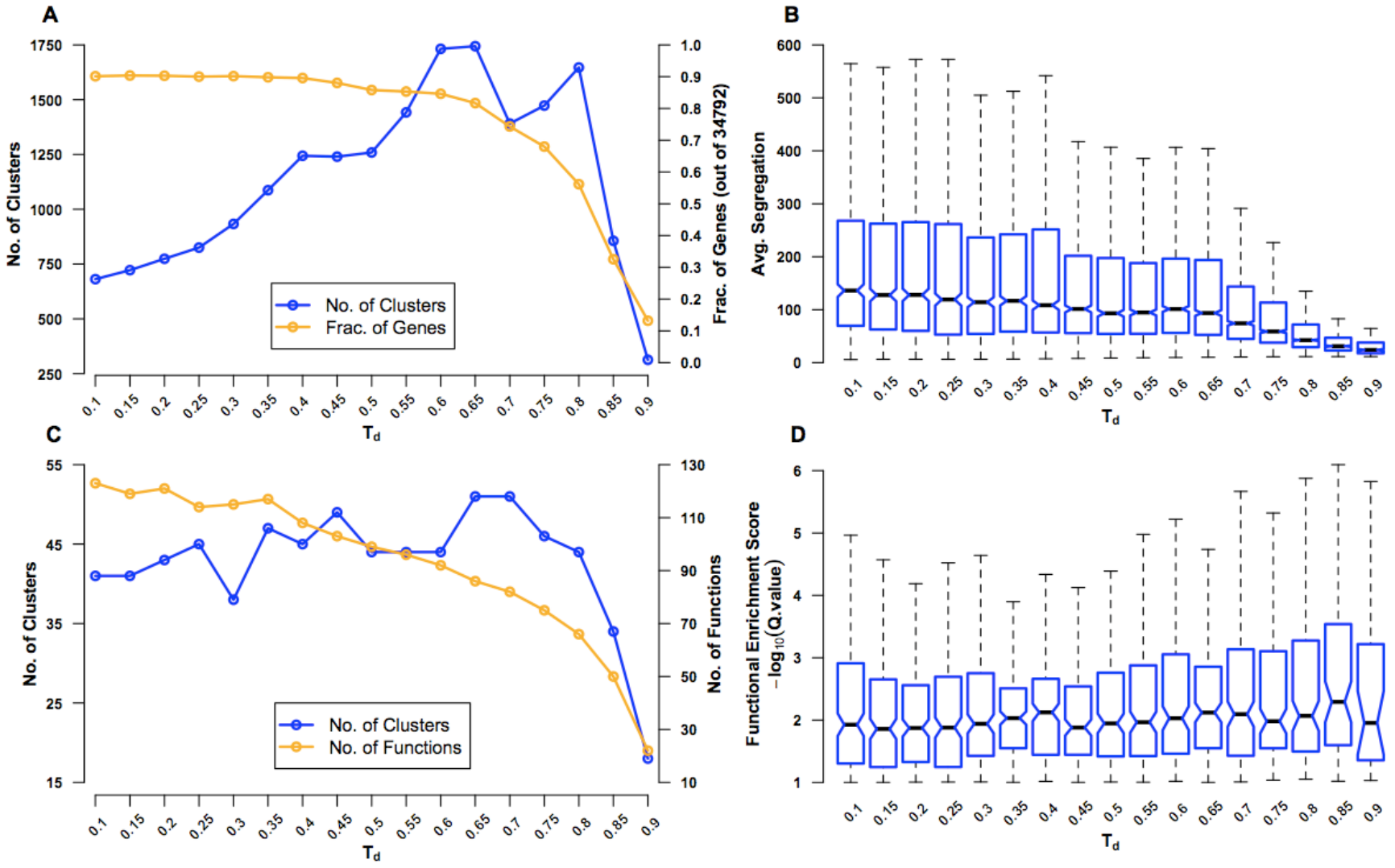

A
No. of Clusters
Frac. of Genes (out of 34792)
No. of Clusters
Frac. of Genes
T_d
B
Avg. Segregation
T_d
C
No. of Clusters
No. of Functions
No. of Clusters
No. of Functions
T_d
D
Functional Enrichment Score
$-\log_{10}$(Q.value)
T_d

Figure 7.5: Evaluation of coexpression network clustering.
The rice 'environment' coexpression network was clustered using SPICi, for a range of values – 0.1-0.9 – of the density parameter T_d that determines how dense the final clusters are. The clusters obtained using each T_d value were evaluated using several criteria: (**A**) Number of clusters that were formed (left y-axis) and the fraction of 34,792 genes in the original network present in one of the clusters (right y-axis) are plotted. These numbers were calculating by considering only clusters containing 3 or more genes. As T_d increases, more and more genes are left out of clusters. (**B**) Average segregation of a cluster is a measure of how well genes in that cluster interact with other genes belonging to the same cluster compared to interactions with genes belonging to other clusters. Hence, average segregation measures cluster modularity. The overall modularity at a given T_d value is plotted a box plot, leaving out outlier values above the whiskers for clarity. The center of the box corresponds to the median (2^{nd} quartile; Q_2) of the distribution of average segregation values of all the clusters, and the extremes of the box correspond to the 1^{st} (Q_1) and 3^{rd} (Q_3) quartiles. The whiskers denote $Q_2 \pm 1.5*IQR$, where IQR is the interquartile range (Q_3-Q_1). The notches in each box extend to +/-1.58 $IQR/\sqrt{n}$ (n being the sample size) (McGill et al, 1978). They are based on asymptotic normality of the median and roughly equal sample sizes for the medians being compared, and are said to be rather insensitive to the underlying distributions of the samples. The notches give roughly a 95% confidence interval for the difference in two medians. (**C**) The extent of overlap between clusters (defined based on a particular T_d value) and GO BP gene sets (termed 'functions') is measured using the hypergeometric test. The number of clusters with significant overlap (FDR q-$value$ <0.1) (left y-axis) and number of distinct functions significantly overlapping with the clusters (right y-axis) are plotted. (**D**) Functional enrichment of the clusters is quantified using $-log_{10}(q\text{-}value)$ and plotted using a box plot representing the distribution of the enrichment scores for all the clusters at a given T_d value. Here again, outliers beyond the whiskers have been left out for clarity.

Second, we calculated a measure of modularity called average segregation that quantified how well genes within a cluster are connected to each other compared to their connection to all the genes in the network (Fig. 7.5B) (Yook et al, 2004). Since we are interested in finding coherent biological modules, finding a T_d that preserves segregation is sought after. It was surprising that the network showed the highest values of segregation for the smallest values of T_d, indicating that even the original network with ~18.5 million edges is highly modular. Therefore, in the context of this network, at least, it was only important to look out for partitioning the network as much as possible without a significant drop in the inherent modularity. The first significant drop in average segregation (measured more qualitatively than quantitatively using the notches in the box plots; see Fig. 7.5 legend) occurs when the T_d value is increased from 0.65 to 0.70, suggesting that setting T_d=0.65 ensures the maximum modularity-preserving partitioning of the network.

Third, as we are interested in the functional consistency of genes within a cluster in addition to topological cohesiveness, we characterized the functional enrichment of all the clusters for a given T_d value using GO BP enrichment analysis (Fig. 7.5C). Since this approach will suffer from the very sparse functional annotation of rice genes, we used this analysis only as a rough guide. Following the number of clusters that were significantly enriched with at least one

specific GO BP ('function'), we observed that the maximum enrichment again occurs at T_d=0.65 (slightly better than T_d=0.70). However, contrary to what is expected, the number of distinct enriched functions dropped steadily with increasing T_d. Finally, using data from the enrichment analysis, we plotted the distribution of enrichment scores of all the clusters for different T_d values and found that T_d values in the range of 0.65 to 0.80 were giving overall more significant overlap between clusters and functions (Fig. 7.5D). Therefore, based on all the four analyses, we decided on a T_d=0.65 to be the best choice for clustering RECoN.

We subsequently clustered RECoN using SPICi with T_d=0.65 to uncover 1744 dense clusters with 3 or more genes. 28,421 genes (~81.7% of all the genes in the original network) fell within one of the clusters. Figure 7.6 shows the diverse expression profiles of clusters of genes across the set of 45 conditions (from the original gene expression data used for coexpression network analysis). Clustering the conditions based on their expression profiles also yields an expected grouping, especially with the drought-, salt- and cold stress samples clustering together.

To find clusters relevant to drought stress, each of the 1744 clusters were tested for enrichment of drought-responsive (up- or down-regulated) genes from any one of the stages (seedling, vegetative or reproductive) determined previously (Fig. 7.4, step 2). Drought clusters provide a handle on putative functional interactions between genes transcriptionally regulated by drought that were otherwise unassociated parts lists. This makes gene-by-gene interpretation a much easier and constructive process. Moreover, we reasoned that since a cluster is a coherent group of genes, all the genes in a 'drought' cluster might have a role in mediating drought-response, not necessarily by responding to drought through gene expression changes. This is possible by either being ubiquitously present as support machinery (between well-watered and drought conditions) or being conditionally active under drought due to non-transcriptional modes of regulation including post-translational modification. These clusters, hence, provide a means for functionally associating post-transcriptionally modified regulatory/signaling genes to transcriptionally regulated genes.

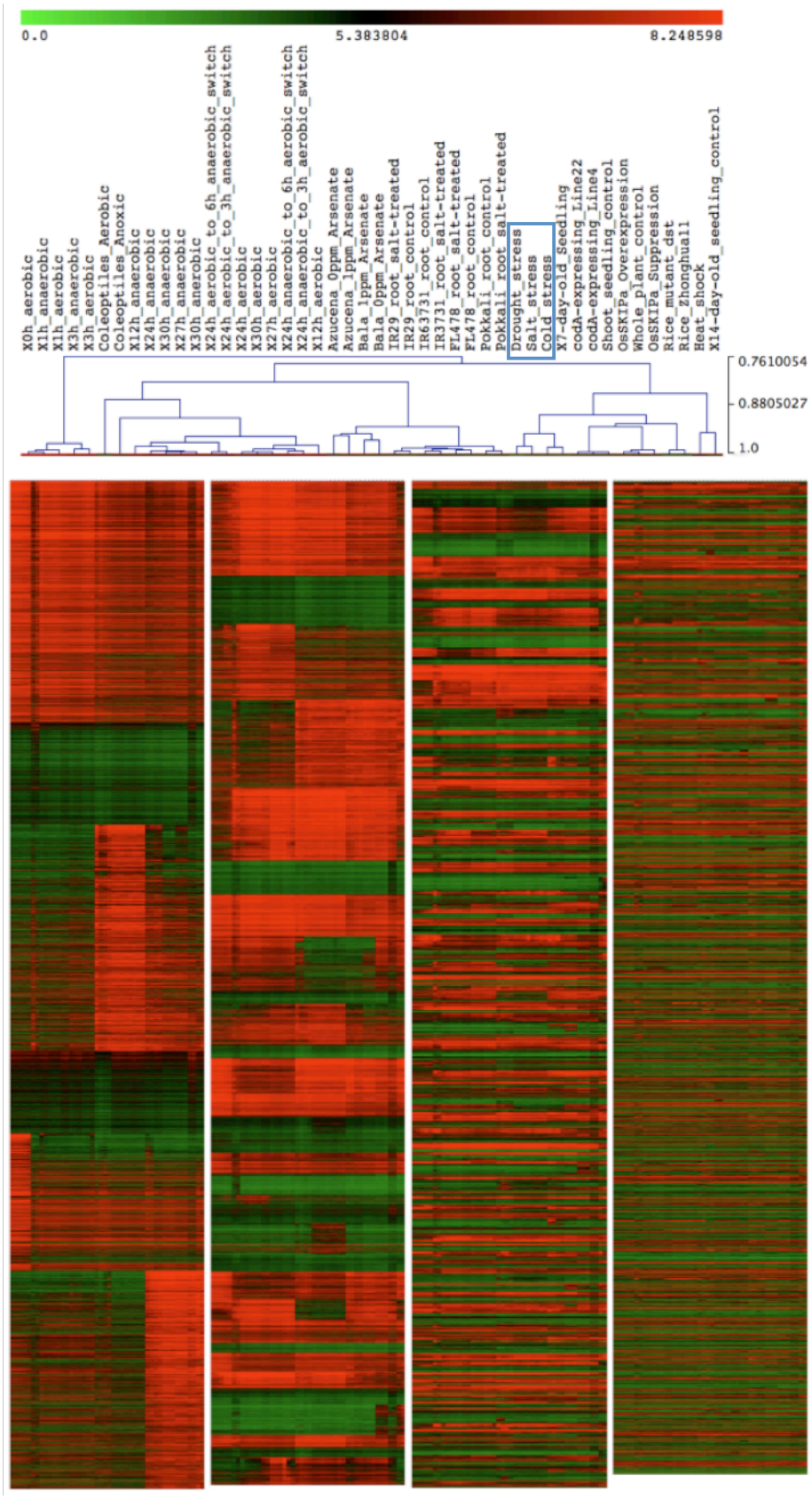

0.0
5.383804
8.248598
X0h_aerobic
X1h_anaerobic
X1h_aerobic
X3h_anaerobic
X3h_aerobic
Coleoptiles_Aerobic
Coleoptiles_Anoxic
X12h_anaerobic
X24h_anaerobic
X30h_anaerobic
X27h_anaerobic
X30h_anerobic
X24h_anaerobic_to_6h_anaerobic_switch
X24h_aerobic_to_3h_anaerobic_switch
X24h_aerobic
X30h_aerobic
X27h_aerobic
X24h_anaerobic_to_6h_aerobic_switch
X24h_anaerobic_to_3h_aerobic_switch
X12h_aerobic
Azucena_0ppm_Arsenate
Azucena_1ppm_Arsenate
Bala_1ppm_Arsenate
Bala_0ppm_Arsenate
IR29_root_salt-treated
IR29_root_control
IR63731_root_control
IR3731_root_salt-treated
FL478_root_salt-treated
FL478_root_control
Pokkali_root_control
Pokkali_root_salt-treated
Drought_stress
Salt_stress
Cold_stress
X7-day-old_Seedling
codA-expressing_Line22
codA-expressing_Line4
Shoot_seedling_control
OsSKIPa_Overexpression
Whole_plant_control
OsSKIPa_Suppression
Rice_mutant_dst
Rice_Zhonghua11
Heat_shock
X14-day-old_seedling_control
0.7610054
0.8805027
1.0

Figure 7.6: Gene expression profiles of the 28,421 genes across the 45 conditions/groups, organized based on coexpression cluster membership of genes.
For ease of visualization, the genes have been split into 4 groups (blocks) of ~7100 genes each. Conditions indicated on top were clustered based on similarity in expression of all the genes (Pearson correlation; Average linkage) and correspond to the columns in each block. Colors indicate the level of expression (in log_2 scale) with red, black and green corresponding to a high, moderate and low expression levels.

For further characterization of the 'drought' clusters thus obtained, for each cluster of genes, we extracted four sets of genes: all the genes in the cluster, and seedling, vegetative and reproductive drought-regulated genes. We then analyzed these gene sets from each cluster for discovery of putative CREs, enrichment of GO biological processes or RiceCyc biochemical pathways, and mapping to known abiotic stress QTL intervals in the rice genome (Fig. 7.4, step 3). In a proposed step, we plan to present all this data to the user via a dynamic visual interface where the user can explore functional annotations and QTL-mapping in any cluster of interest to come up with a small set of candidate genes (involved in drought response and possibly in drought tolerance) for testing (Fig. 7.4, step 4). This step requires a flexible network exploration tool that overlays different combinations of annotations on a cluster that will aid in finding nonobvious/novel candidate genes using some of the following criteria: a) present in a 'drought' cluster; b) connected significantly to 'known' drought response/tolerance genes; c) contains abiotic stress-associated CREs in their upstream regions; d) part of a biological process or biochemical pathway that might be involved in drought or other abiotic stress response; e) mapping within or very close to known abiotic stress QTL intervals. Finally, in the last step, the genes identified by the user in the previous step can be parsed through a rice mutant resource compendium (Krishnan et al, 2009) to obtain mutants of the candidate genes that can be obtained for testing in the laboratory.

7.3.4. Examples of drought transcriptional modules

We have presented here some drought transcriptional modules here as examples to showcase the usefulness of this approach in understanding developmental stage-specific drought response. All the genes in drought clusters, their relative expression across the stages and their 'drought' cluster membership are provided in Table 7.S3.

Cluster0013 contains 294 genes enriched with genes up-regulated in the seedling stage and down-regulated in the reproductive stage. Genes in this cluster are involved in ribosome

biogenesis and mitochondrial protein localization (which concerns transporting of nuclear genome-coded mitochondrial oxidative phosphorylation proteins to the mitochondrion), and contain the GCC-core, Telo-box and the Site II motifs in their upstream sequences. This combination of biological processes and CREs represents a well known regulatory program: the site II motifs are recognized by TFs of the TCP family and have been confirmed to be important in the regulation of ribosome protein (RP) genes in combination with the telo-box motif (Tremousaygue et al, 2003). These motifs are co-located in the promoters of about 70% of 216 ribosomal protein genes in Arabidopsis. In addition, there is evidence that the site II motifs also possibly coordinate the expression of nuclear genes encoding components of the mitochondrial oxidative phosphorylation machinery in both Arabidopsis and rice (Welchen & Gonzalez, 2006). Therefore, this program involving site II and telo-box motifs could mediate the down-regulation of major processes that affect protein production under drought stress in the reproductive tissue. The GCC-core motif is known to be bound by AP2-ERF TFs (Ohme-Takagi & Shinshi, 1995), which are involved in gene regulation under a variety of abiotic stresses conserved between Arabidopsis and rice (Nakashima et al, 2009).

Cluster0010 contains 635 genes including genes involved in lignan biosynthetic process, amino acid transport, systemic acquired resistance, glycolysis, pentose-phosphate shunt and two-component signal transduction system (phosphorelay). Genes in this cluster are down-regulated in the seedling and vegetative stages, but up-regulated in reproductive stage. Of particular interest in this cluster is the OsVIN1 gene (LOC_Os04g45290) coding for a vacuolar invertase gene. Invertases play an important role in carbon allocation to developing organs like the reproductive tissue. OsVIN1 is not induced by our drought treatment, and this is in agreement with the observation that OsVIN1 is expressed in flag leaves, panicles (the reproductive tissue) and anthers in an essentially drought-insensitive manner (Ji et al, 2005b). It is therefore a case where a gene involved in mediating a process (resource allocation) relevant to drought is not transcriptionally affected, but is associated with other drought-regulated genes in clusters defined by us.

Cluster	No. Genes	No. Dr.Tol. Genes	GeneID:GeneName	Rice Dr		
				Seedl'g	Veg	Rep
Cluster0079	71	4	Os01g07120:**OsDREB2A** | Os05g46480:**OsLea3-1** | Os05g51670:**OsUGE-1** | Os01g66120:**SNAC2/OsNAC6**	24.829	28.811	13.734
Cluster0001	1713	3	Os05g28350:**OsABI4** | Os07g01770:**OsSMCP1** | Os08g06050:**OsTOP6A2/SPO11-2**	196.385	4.990	0.000
Cluster0424	20	3	Os01g73770:**OsDREB1F** | Os01g50400:**OsNPKL2** | Os01g50410:**OsNPKL3**	0.000	0.000	1.509
Cluster0005	972	2	Os01g64000:**OsABI5** | Os08g45110:**OsDREB2C**	134.490	93.440	13.338
Cluster0006	1545	2	Os03g57240:**DST** | Os01g03570:**OXHS2**	-296.778	-197.347	-85.746
Cluster0018	252	2	Os02g52780:**OsbZIP23** | Os09g28310:**OsbZIP72**	82.263	79.406	8.123
Cluster0050	34	2	Os01g58420:**AP37** | Os05g38660:**OsSIK1**	7.315	6.867	1.816
Cluster0051	62	2	Os11g03300:**OsNAC10** | Os11g08210:**OsNAC5**	6.671	9.003	0.000
Cluster0119	52	2	Os09g35030:**OsDREB1A** | Os09g35010:**OsDREB1B**	12.484	6.681	1.584
Cluster0177	33	2	Os03g17700:**OsMAPK5** | Os01g50420:**OsNPKL4**	2.211	2.375	0.000
Cluster0216	15	2	Os02g43790:**AP59** | Os12g39400:**ZFP252**	1.937	3.514	0.000
Cluster0003	1601	1	Os01g50370:**OsNPKL1**	-63.469	4.752	12.491
Cluster0011	507	1	Os02g50970:**DSM1**	-112.341	-4.211	-10.141
Cluster0014	392	1	Os01g55450:**OsCIPK12**	8.685	-18.297	0.000
Cluster0016	290	1	Os07g48760:**OsCIPK03**	-1.248	2.201	0.000
Cluster0019	160	1	Os06g41010:**OsSAP8**	24.531	22.722	0.000
Cluster0049	75	1	Os05g27930:**OsDREB2B**	21.794	11.431	0.000
Cluster0062	65	1	Os01g01420:**OsCOIN**	2.680	-4.680	0.000
Cluster0072	38	1	Os05g45020:**OsMT1a**	5.377	6.678	1.198
Cluster0108	37	1	Os03g08310:**OsTIFY11a/OsJAZ9**	1.179	0.000	1.700
Cluster0111	29	1	Os02g50930:**OsBIRF1**	0.000	2.342	0.000
Cluster0187	23	1	Os01g62410:**OsMYB3R-2**	10.406	9.990	0.000
Cluster0218	13	1	Os03g17610:**OsTOP6A3/SPO11-3**	1.249	0.000	0.000
Cluster0352	17	1	Os09g10770:**OsTOP6B**	5.451	0.000	0.000
Cluster0396	9	1	Os02g52250:**OsSKIPa**	2.868	1.853	0.000
Cluster0508	14	1	Os04g49510:**OsCDPK7**	2.161	0.000	1.878
Cluster0728	7	1	Os03g60080:**SNAC1**	1.828	1.584	0.000
Cluster1207	4	1	Os02g08440:**OsABF2**	2.212	0.000	0.000
Cluster1478	4	1	Os02g43970:**ARAG1**	-1.285	0.000	0.000

Table 7.1: Drought clusters containing known drought tolerance genes.
Genes in black are regulated by drought at one or more of the growth stages while genes in red are not drought-regulated. The values in the color-coded columns correspond to the level of significance (measured as score equal to the $-log_{10}[q\text{-}value]$) of drought-regulated genes. For convenience the scores themselves are signed and colored based on the direction of their regulation (+/blue - up-regulation; –/yellow - down). Since only enrichments with *q-value*<0.1 were considered, all the other values were set to 1 (because of which, their negative logarithms are 0s).

The 193 genes in Cluster0041 are enriched primarily in almost all processes involved in cell cycle, a process integral to panicle development and elongation, and these genes are specifically down-regulated by drought at the reproductive stage (the most drought sensitive stage of rice). Upstream regions of these genes contain the SEF3 binding site/ACII element, MYB recognition site found in *rd22* and other genes, and E2F consensus, potential binding sites of TFs that have been implicated to be important in regulating cell cycle in the reproductive tissue of Arabidopsis (Hennig et al, 2004).

The other aspect of the application of this approach is in discovery drought tolerance gene discovery. A variety of gene families with regulatory function have been shown to impart drought tolerance by overexpression/knockout and regulation of a battery of downstream genes (Umezawa et al, 2006). Therefore, to evaluate this aspect, we first catalogued a number of genes that confer drought tolerance in rice on overexpression or knockout, and then mapped them to RECoN clusters (Table 7.1). The primary observation is that almost all the drought tolerance genes were part of drought clusters. However, this observation could be trivial if all those genes are indeed regulated by drought in the first place. Out of the 44 genes presented here, while 34 are indeed regulated by drought in one stage or the other, 10 of these are not drought-regulated, but are associated with a drought module. Therefore, we reaffirm that the approach lends itself to identification of genes that are not necessarily transcriptionally perturbed by drought, if at all regulated by it. Some examples for the drought-tolerance clusters follow.

Cluster0079 contains 71 genes including four tolerance genes OsDREB2A (LOC_Os01g07120), OsLea3-1 (LOC_Os05g46480), OsUGE-1 (LOC_Os05g51670) and SNAC2/OsNAC6 (LOC_Os01g66120) (Fig. 7.7). Most genes in the cluster including the four tolerance genes are up-regulated by drought in all stages and accordingly have a ABRE-like motif in their upstream sequences. While everything about this cluster points to the fact that it is a typical drought-stress response module, it contains several uncharacterized genes. These genes can be candidates for experimental verification.

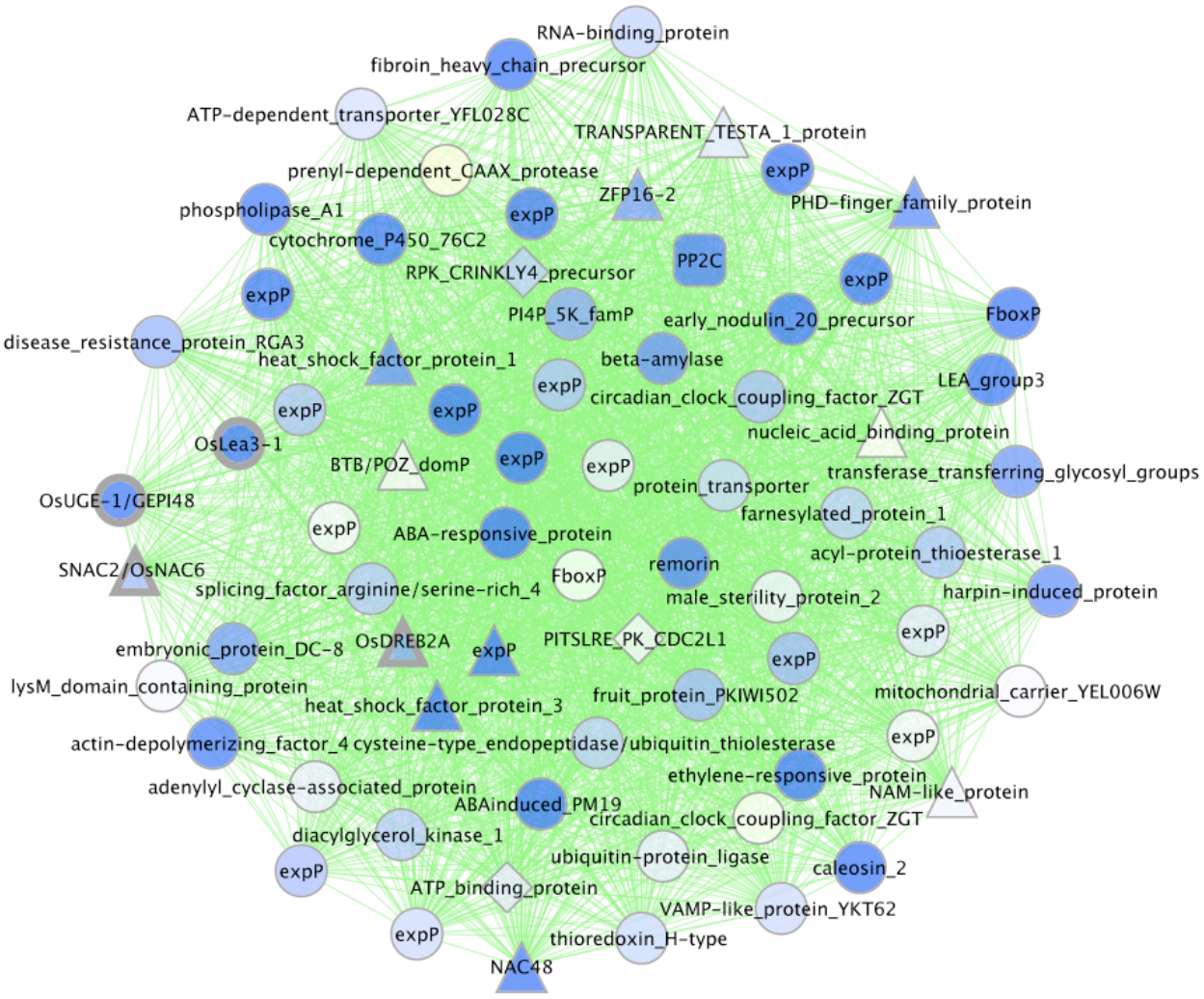


Figure 7.7: 71 genes in Cluster0079 that contains four drought tolerance genes (with thick grey borders). All the coexpression edges are colored green. Node shapes correspond to type of gene: triangles are TFs, diamonds are protein kinases, rounded squares are protein phosphatases and circles are other genes. Node color corresponds to the level of differential expression under drought in the vegetative stage (where this cluster has maximum enrichment): blue for up-regulation and yellow for down-regulation. Most genes in this cluster are uncharacterized (labeled 'expP' for 'expressed protein' or 'hypP' for 'hypothetical protein').

Cluster0424 contains 20 genes enriched specifically with reproductive drought-regulated genes and these genes too contain an ABRE-like motif – HACGYGTNS – in their upstream sequences. Drought tolerance genes OsDREB1F (LOC_Os01g73770), OsNPKL2 (LOC_Os01g50400) and OsNPKL3 (LOC_Os01g50410), where NPKL2 and NPKL3 are previously known to be strongly induced by drought precisely at the reproductive stage (Ning et al, 2008). In a previous study, OsDREB1F was induced by abiotic stresses including drought (using PEG) (Wang et al, 2008). However, our progressive drought treatment does not perturb this gene (at least not at the stringent level of significance chosen). Developmental stage-specific drought-regulation of

OsDREB1F is not clear except that the gene by itself is expressed differently in different stages and tissues. We therefore implicate OsDREB1F as being important in progressive drought response in the reproductive stage.

Cluster0177 contains 33 genes involved in the regulation of innate immune/defense/stress response as well as response to jasmonic acid and salicylic acid. Drought regulated genes in this cluster are up-regulated specifically in the seedling and vegetative stages. This cluster again contains the drought-tolerance genes OsMAPK5 (LOC_Os03g17700) and OsNPKL4 (LOC_Os01g50420). OsMAPK5 is known to be induced by drought, other abiotic stresses and ABA, as well as pathogen infection (Xiong & Yang, 2003). It is hence considered to be a key link in the cross talk between disease resistance and abiotic stress tolerance. We propose that other genes in this cluster are links between the abiotic-biotic stresses. Previous work has observed that, under drought, OsNPKL4 showed very strong induction at the seedling stage but only a moderate or low level of induction at the anthesis stage (Ning et al, 2008), consistent with the drought-pattern of this cluster.

7.4. Discussion

Plant response to environmental stress is cut across several layers of organization including signaling, transcription and metabolism, making it vital to understand stress response at the systems-level. For less studied models like rice, the current scope for systems analysis is mostly restricted to transcriptional profiling under various conditions. Therefore, to make the best use of currently available data in rice, we have created a resource for exploration of transcriptional, developmental, functional, regulatory and genomic aspects of drought response in rice.

Using gene expression profiling of drought treated rice plants at the seedling, vegetative and reproductive stages, we identified that a large number of genes are perturbed in all stages. We performed functional enrichment analysis and observed that processes that are characteristic of general drought response are altered differently in the three stages suggesting developmental stage specific responses. *De novo* CRE discovery showed that the ABRE-like motif is the only regulatory element that is involved in drought induction of gene expression in all three stages. Even while photosynthesis is repressed in all stages, the CREs mediating this repression appear

to be different in the three stages. Moreover, as a result of this analysis, at least four distinct novel motifs putatively involved in specific drought responses were identified. However, rice genes are extremely poorly annotated with function, so that gleaning any more insight into transcriptional regulation of specific processes is difficult.

Therefore, to make use of other available data, we sought to organize drought response genes into coherent groups and work further from there. To this end, we deigned and implemented a pipeline for automatic mining of condition-specific gene expression datasets intended for analysis of coexpression. At a practical level, accurate quantification of gene expression using technologies like Affymetrix GeneChips has been hard due to the problem of cross-hybridization. This has been noted to affect calculation of coexpression (Casneuf et al, 2007) and the proposed solution is a remapping of microarray probes to genes to ensure unique hybridization (Dai et al, 2005). We hence created a custom probe-gene mapping and used this reannotation to make reliable estimation of gene expression across 45 conditions. Then, a coexpression network was built (RECoN) and clustered to obtain tightly coexpressed groups of genes that revealed the modular organization of genes.

Coexpression networks (or co-responsive genes in a simple gene expression study) are primarily pointing to similarity in changes in gene expression that could arise due to similar transcriptional or post-transcriptional regulation of the amount of mRNA of the genes. Therefore, coexpression clusters/modules could be used not only to understand overall organization of changes in gene expression, but also to study transcription factor-mediated or small RNA (e.g. miRNA) mediated regulation. As a first pass, this could be carried out by testing if predicted targets of a TF or a miRNA are enriched in a module. Moreover, the *de novo* CRE analysis can be similarly extended to identification of DNA sequences in the 3' UTR of genes that are informative about the expression pattern of the genes.

In this work we focus on understanding drought response and discover novel drought tolerance genes at the organizational level. For this, we combined drought-responsive genes from our experiments with the transcriptional modules to uncover drought clusters, where each cluster, by design, contains several other genes in addition to genes transcriptionally regulated by drought.

Drought modules thus present an opportunity to discover regulatory genes that do not change in gene expression and affect the response mediated by that module. In this process, we are basically imputing uncharacterized genes within a cluster with the function/role of characterized genes (even at the level of transcriptional response). In species with very little annotation, such as rice, cluster-level function prediction has been shown to be useful (Song & Singh, 2009). We have validated this approach by inspecting the cluster membership of known drought tolerance genes that are not drought responsive but are associated with a cluster that is enriched in genes following a drought expression pattern expected from what is known about the tolerance gene.

With enormous amount of data generated in this work that can be used for inference, in the future work, all these results will be summarized and presented in a flexible visual interface for dynamic exploration. The approach presented here is widely applicable: genome-wide transcriptional modules recovered here on the basis of gene expression under different environmental conditions can be similarly extended to study other abiotic stresses including salt and cold to find common stress-specific modules.

7.5. Methods

7.5.1. Gene Expression Analysis

Total RNA was isolated from the drought treated rice japonica cv. Nipponbare at the seedling (2 weeks), vegetative (V4) and reproductive (R4) stages (Counce et al, 2000) of wild-type plants (along with well-watered controls) and was used for hybridization to rice Affymetrix GeneChips.

7.5.2. Reannotation of Rice GeneChip Probe-Gene Mapping

A high-quality custom chip definition file (CDF) was built for the rice GeneChip array by uniquely mapping 442,810 probe sequences (http://www.affymetrix.com/analysis/downloads/data/) to 35,161 rice gene-based probe sets in the following manner: (i) probes that have perfect sequence identity with a single target gene were selected, (ii) probes mapping to reverse complements of genes were annotated separately as antisense probes (not used in the above counts), and finally, (iii) probes were grouped into probe sets, each corresponding to a single gene, and probe sets with at least 3 probes were retained (>98% probe sets have ≥5 probes). Note that these stringent criteria used to construct the CDF

make it possible to reliably measure expression values of members of multigene families (free from cross-hybridization between paralogs showing high sequence similarity) and to get around 'one gene to multiple probe sets' ambiguities. Previous work in humans and other systems has shown that such reannotation procedures significantly alter the interpretation of gene expression measured using GeneChips (Dai et al, 2005).

7.5.3. Analysis of Differential Gene Expression

Raw data from all the drought experiments were background corrected, normalized and summarized according to the custom CDF using RMA (Gentleman et al, 2004; Ihaka & Gentleman, 1996; Irizarry et al, 2003), followed by non-specific filtering of genes that do not have enough variation (interquartile range (IQR) across samples < IQR_{median}) to allow reliable detection of differential expression. A linear model was then used to detect differential expression of the remaining genes (Smyth, 2004). The *p*-values from the moderated *t*-tests were converted to *q*-values to correct for multiple hypothesis testing (Storey & Tibshirani, 2003), and genes with *q*-value <0.01 were declared as differentially expressed in response to drought.

7.5.4. Coexpression Network Analysis

29 publicly available Affymetrix rice GeneChip gene expression datasets (414 samples; 150 groups after gathering biological replicate samples into single groups) were collected from NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/) (Barrett et al, 2009) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) (Parkinson et al, 2009), and the largest subset of experiments (10 datasets; 129 samples; 45 groups) with a similar biological context corresponding to studies of response to some environmental condition was used for coexpression analysis (see Table 7.S4).

Raw data were background corrected, normalized and summarized according to the custom CDF using justRMA (Irizarry et al, 2003), and expression values were averaged across replicates. Pearson correlations were first calculated between every pair of genes (Huttenhower et al, 2008), which were then Fisher Z-transformed (David, 1949) and standardized to get coexpression scores (z_{cs}) with a N(0,1) distribution. This formulation was robust and highly interpretable as deviations from the expected value, and even by level of significance where $|z_{cs}|$ values greater

than 1.645, 1.96 and 2.58, for example, correspond to 10%, 5% and 1% extremes of the distribution of z_{cs} scores.

A coexpression network was then constructed connecting pairs of genes that have a z_{cs} >1.96 (top 2.5% of all pairs of genes ordered in decreasing order of correlation). This cutoff corresponded to a Pearson correlation coefficient of 0.632. This network contains 34,792 genes connected by ~18.5 million edges was then clustered using SPICi (Jiang & Singh, 2010). Since SPICi requires a density parameter T_d as input, a range of values of the parameter from 0.1 to 0.9 was tested. Clusters obtained using each T_d value were evaluated using several criteria including the number of clusters formed, fraction of genes in clusters of size 3 or more, average segregation (modularity), and extent of overlap between clusters and GO BP gene sets (termed 'functions') (Fig. 7.4). In order to calculate average segregation, as desired property of dense interaction networks, the coexpression network is modeled as an undirected graph $G=(V, E)$, consisting of a set V of nodes (i.e., genes) and a set E of edges (i.e., coexpressing gene pairs). Let w_{uv} denote the weight of the edge $(u, v) \in E$, denoting the Pearson correlation coefficient of gene pairs (u, v). The graph $G_c=(V_c, E_c)$ is defined as the graph induced by the genes that are part of cluster c, and average segregation is computed as:

$$\frac{\sum_{(u,v)\in E_c} w_{uv} \Big/ |E_c|}{\sum_{u\in V_c; v\in V; (u,v)\in E} w_{uv} \Big/ |E_{c'}|}$$

where $E_{c'}$ is the set of edges in G that are incident on V_c. For functional enrichment analysis the overlap between genes within a cluster and genes annotated to a given GO BP term using the cumulative hypergeometric test. Using only GO BP terms that annotate <500 genes (to ensure a certain level of specificity in definition), for a pair of gene sets (cluster and GO BP term) i and j, if N is the total number of genes, n_i and n_j are the number of genes in gene set i and j, and m is the number of genes common to the gene sets, the probability (p-value) of an overlap (enrichment) of size equal to or greater than observed is given by the formula below.

$$P(X = x \geq m) = \sum_{x=m}^{\min(n_i,n_j)} \frac{\binom{n_i}{x}\binom{N-n_i}{n_j-x}}{\binom{N}{n_j}}$$

P-values from the test were converted to *q*-values to correct for multiple hypothesis testing using Benjamini-Hochberg method (Benjamini & Hochberg, 1995) and cluster-GO_BP pairs with *q*-value <0.1 were considered for analysis. The level of functional enrichment in a cluster is quantified using $-log_{10}$(*q*-value).

After clustering the network using SPICi with T_d value of 0.65, from all the clusters, those relevant to drought were determined by testing which clusters contained a significantly high number of drought-regulated genes up- or down-regulated in any one of the stages (again using a cumulative hypergeometric test). Then, four sets of genes were extracted from each 'drought' cluster – all the genes in the cluster, and seedling, vegetative and reproductive drought-regulated genes – for discovery of putative cis-regulatory elements, enrichment analysis of GO biological processes or RiceCyc biochemical pathways, mapping to known abiotic stress QTL intervals in the rice genome.

7.5.5. Promoter analysis and CRE discovery

For analysis of potential promoter-resident cis-regulatory elements (CREs), FIRE (Elemento et al, 2007) was used to discover motifs informative about the different sets of differentially expressed genes (e.g. SeedlingDr_Up or Rep_Dr_Down) or clusters of genes compared to the rest of the genes in the genome. Briefly, FIRE seeks to discover motifs whose patterns of presence/absence across all considered regulatory regions (motif profile) are most informative about the expression of the corresponding genes (expression profile). To measure these associations, FIRE uses mutual information (MI) (Cover & Thomas, 2006). FIRE performs a randomization test and considers an observed MI value (for a motif-expression profile pair) to be significant only when it is greater than all the random MI values calculated by randomly assigning the expression values to genes. A Z-score reflecting how far the observed MI value is, in number of standard deviations, from the average random MI is calculated. These are the Z-scores presented for each motif in Figure 7.3. Moreover, it also performs jack-knife resampling (Efron, 1979), where, in each of 10 trial the above randomization test is carried out. Only motifs that are statistically significant in at least 6 trials are reported. Newly discovered motifs were compared to known cis-elements in the PLACE database (Higo et al, 1999) and to each other using STAMP (Mahony & Benos, 2007). All upstream sequences were obtained from the rice

genome annotation database (Ouyang et al, 2007). This *de-novo* approach was taken since i) CREs could diverge far more quickly than coding sequences across species, making them hard to find simply by searching, and ii) searching based on known elements is limited by the scope of experimental identification in a select set of genes, making identification of degenerate yet potentially functional positions in the element hard.

7.5.6. Functional annotation of drought clusters

GO BP process and RiecCyc pathway enrichment analysis was performed using the cumulative hypergeometric test (as described above). RiceCyc pathway annotations were downloaded from Gramene (Liang et al, 2008). GO BP annotations for rice genes were obtained from the EasyGO webserver (Zhou & Su, 2007). Since the GO annotations in rice are extemenly spare, to annotate drought clusters, GO annotations of genes were borrowed from their Arabidospis homologs. Homology was defined based on InParanoid (Remm et al, 2001) and Arabidopsis annotations were obtained from TAIR (Swarbreck et al, 2008).

Chromosomal positions of abiotic stress QTLs were obtained from Gramene (Liang et al, 2008) and each locus in the rice genome was mapped to one of the 336 QTL intervals if it's mid-point ([start-coordinate + stop-coordinate]/2) fell within that interval. Mutants of candidate genes is made available to this pipeline from a rice mutant resource compendium (Krishnan et al, 2009).

Perl scripts were used to parse all the data. Plots were generated using R (Ihaka & Gentleman, 1996) and gene expression matrices were visualized using MeV (Saeed et al, 2006).

7.6. References

Aharoni A, Dixit S, Jetter R, Thoenes E, van Arkel G, Pereira A (2004) The SHINE clade of AP2 domain transcription factors activates wax biosynthesis, alters cuticle properties, and confers drought tolerance when overexpressed in Arabidopsis. *Plant Cell* 16: 2463-2480

Andersen MN, Asch F, Wu Y, Jensen CR, Naested H, Mogensen VO, Koch KE (2002) Soluble invertase expression is an early target of drought stress during the critical, abortion-sensitive phase of young ovary development in maize. *Plant Physiol* 130: 591-604

Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4: 2

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* 37: D885-890

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 57: 289-300

Boyer JS (1982) Plant productivity and environment. *Science* 218: 443-448

Casneuf T, Van de Peer Y, Huber W (2007) In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics* 8: 461

Chan CS, Guo L, Shih MC (2001) Promoter analysis of the nuclear gene encoding the chloroplast glyceraldehyde-3-phosphate dehydrogenase B subunit of Arabidopsis thaliana. *Plant Mol Biol* 46: 131-141

Counce PA, Keisling TC, Mitchell AJ (2000) A uniform, objective, and adaptive system for expressing rice development. *Crop Sci* 40: 436-443

Cover TM, Thomas JA (2006) *Elements of information theory*, 2nd edn. Hoboken, N.J.: Wiley-Interscience.

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 33: e175

David FN (1949) The moments of the Z and F distributions. *Biometrika* 36: 394–403

Deyholos MK (2010) Making the most of drought and salinity transcriptomics. *Plant Cell Environ* 33: 648-654

Dinneny JR, Long TA, Wang JY, Jung JW, Mace D, Pointer S, Barron C, Brady SM, Schiefelbein J, Benfey PN (2008) Cell identity mediates the response of Arabidopsis roots to abiotic stress. *Science (New York, NY)* 320: 942-945

Efron BL (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7: 1-26

Elemento O, Slonim N, Tavazoie S (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* 28: 337-350

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30: 1575-1584

Ficklin SP, Luo F, Feltus FA (2010) The Association of Multiple Interacting Genes with Specific Phenotypes In Rice (Oryza sativa) Using Gene Co-Expression Networks. *Plant Physiol*

Fu FF, Xue HW (2010) Co-expression analysis identifies Rice Starch Regulator1 (RSR1), a rice AP2/EREBP family transcription factor, as a novel rice starch biosynthesis regulator. *Plant Physiol*

Fukushima A, Kanaya S, Arita M (2009) Characterizing gene coexpression modules in Oryza sativa based on a graph-clustering approach. *Plant Biotechnol* 26: 485-493

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge YC, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M,

Rossini AJ, Sawitzki G et al (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 5: -

Giuliano G, Pichersky E, Malik VS, Timko MP, Scolnik PA, Cashmore AR (1988) An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. *Proc Natl Acad Sci U S A* 85: 7089-7093

Hartwell L, Hopfield J, Leibler S, Murray A (1999) From molecular to modular cell biology. *Nature* 402: C47-C52

Hattori T, Totsuka M, Hobo T, Kagaya Y, Yamamoto-Toyoda A (2002) Experimentally determined sequence requirement of ACGT-containing abscisic acid response element. *Plant Cell Physiol* 43: 136-140

Hennig L, Gruissem W, Grossniklaus U, Kohler C (2004) Transcriptional programs of early reproductive stages in Arabidopsis. *Plant Physiol* 135: 1765-1775

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

Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG (2008) The Sleipnir library for computational functional genomics. *Bioinformatics* 24: 1559-1561

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. *J Comput Graph Stat* 5: 299-314

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003) Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Research* 31: -

Ji X, Van den Ende W, Van Laere A, Cheng S, Bennett J (2005a) Structure, evolution, and expression of the two invertase gene families of rice. *J Mol Evol* 60: 615-634

Ji XM, Raveendran M, Oane R, Ismail A, Lafitte R, Bruskiewich R, Cheng SH, Bennett J (2005b) Tissue-specific expression and drought responsiveness of cell-wall invertase genes of rice at flowering. *Plant Mol Biol* 59: 945-964

Jiang P, Singh M (2010) SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics (Oxford, England)* 26: 1105-1111

Jiao Y, Ma L, Strickland E, Deng XW (2005) Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis. *Plant Cell* 17: 3239-3256

Kasuga M, Liu Q, Miura S, Yamaguchi-Shinozaki K, Shinozaki K (1999) Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor. *Nature Biotechnology* 17: 287-291

Kostka D, Spang R (2004) Finding disease specific alterations in the co-expression of genes. *Bioinformatics* 20 Suppl 1: i194-199

Krishnan A, Guiderdoni E, An G, Hsing YI, Han CD, Lee MC, Yu SM, Upadhyaya N, Ramachandran S, Zhang Q, Sundaresan V, Hirochika H, Leung H, Pereira A (2009) Mutant resources in rice for functional genomics of the grasses. *Plant Physiol* 149: 165-170

Lee TH, Kim YK, Pham TT, Song SI, Kim JK, Kang KY, An G, Jung KH, Galbraith DW, Kim M, Yoon UH, Nahm BH (2009) RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of coexpressed genes in rice. *Plant Physiol* 151: 16-33

Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, Ravenscroft D, Ren L, Spooner W, Tecle I, Thomason J, Tung CW, Wei X, Yap I,

Youens-Clark K, Ware D et al (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res* 36: D947-953

Liu JX, Liao DQ, Oane R, Estenor L, Yang XE, Li ZC, Bennett J (2006) Genetic variation in the sensitivity of anther dehiscence to drought stress in rice. *Field Crop Res* 97: 87-100

Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 35: W253-258

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

Mazzucotelli E, Mastrangelo AA, Crosatti C, Guerra D, Stanca AM, Cattivelli L (2008) Abiotic stress response in plants: When post-transcriptional and post-translational regulations control transcription. *Plant Sci* 174: 420-431

McGill R, Tukey J, Larsen WA (1978) Variations of box plots. *The American Statistician* 32: 12-16

Mentzen WI, Wurtele ES (2008) Regulon organization of Arabidopsis. *BMC Plant Biol* 8: 99

Moore JP, Vicre-Gibouin M, Farrant JM, Driouich A (2008) Adaptations of higher plant cell walls to water loss: drought vs desiccation. *Physiol Plant* 134: 237-245

Nakashima K, Ito Y, Yamaguchi-Shinozaki K (2009) Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. *Plant Physiology* 149: 88-95

Ning J, Liu S, Hu H, Xiong L (2008) Systematic analysis of NPK1-like genes in rice reveals a stress-inducible gene cluster co-localized with a quantitative trait locus of drought resistance. *Mol Genet Genomics* 280: 535-546

Ogata Y, Suzuki H, Sakurai N, Shibata D (2010) CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics* 26: 1267-1268

Ohme-Takagi M, Shinshi H (1995) Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell* 7: 173-182

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res* 35: D883-887

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ et al (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res* 37: D868-872

Rawat A, Seifert GJ, Deng Y (2008) Novel implementation of conditional co-regulation by graph theory to derive co-expressed genes from microarray data. *BMC Bioinformatics* 9 Suppl 9: S7

Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314: 1041-1052

Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J (2006) TM4 microarray software suite. *Methods Enzymol* 411: 134-193

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

Song J, Singh M (2009) How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics (Oxford, England)* 25: 3143-3150

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100: 9440-9445

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

Tremousaygue D, Garnier L, Bardet C, Dabos P, Herve C, Lescure B (2003) Internal telomeric repeats and 'TCP domain' protein-binding sites co-operate to regulate gene expression in Arabidopsis thaliana cycling cells. *Plant J* 33: 957-966

Umezawa T, Fujita M, Fujita Y, Yamaguchi-Shinozaki K, Shinozaki K (2006) Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future. *Curr Opin Biotech* 17: 113-122

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ* 32: 1633-1651

Wang Q, Guan Y, Wu Y, Chen H, Chen F, Chu C (2008) Overexpression of a rice OsDREB1F gene increases salt, drought, and low temperature tolerance in both Arabidopsis and rice. *Plant Mol Biol* 67: 589-602

Wang X, Haberer G, Mayer KF (2009) Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. *BMC Genomics* 10: 284

Welchen E, Gonzalez DH (2006) Overrepresentation of elements recognized by TCP-domain transcription factors in the upstream regions of nuclear genes encoding components of the mitochondrial oxidative phosphorylation Machinery. *Plant Physiol* 141: 540-545

Xiong L, Yang Y (2003) Disease resistance and abiotic stress tolerance in rice are inversely modulated by an abscisic acid-inducible mitogen-activated protein kinase. *Plant Cell* 15: 745-759

Yook SH, Oltvai ZN, Barabasi AL (2004) Functional and topological characterization of protein interaction networks. *Proteomics* 4: 928-942

Zhou X, Su Z (2007) EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics* 8: 246

8. Conclusions

Plants are complex organisms that have evolved sophisticated molecular means to cope with environmental stresses like drought. Understanding the machinery that underlies this response is critical for modifying crop plants to make them tolerant to imminent stresses due to climate change and water resource scarcity. We have used functional genomic tools to capture drought response in Arabidopsis and rice. Using a variety of highly integrative approaches, we have, in each case, unraveled the stress response at various levels of biological organization, from individual genes/proteins, functionally coherent modules, and large-scale networks. In the process, we have also explored fundamental challenges these approaches, including 'gene function prediction' and availability of mutant resources that facilitate experimental validation of genes' predicted roles in stress response. Several integrative analysis frameworks have been presented here that could be used widely for i) detailed mining of regulatory programs connecting regulatory elements to transcriptional regulation of specific biological processes, ii) network-based gene function prediction, iii) elucidation of transcriptional regulatory networks of biochemical pathway genes, iv) exploration of systems-level functional, regulatory and genomic knowledge concerning any abiotic stress response.