12-2017

Transcriptome-based Gene Networks for Systems-level Analysis of Plant Gene Functions

Chirag Gupta
University of Arkansas, Fayetteville

Transcriptome-based Gene Networks for Systems-level Analysis of Plant Gene Functions

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Cell and Molecular Biology

by

Chirag Gupta
Sardar Patel University
Bachelor of Science in Bioinformatics, 2007
Sardar Patel University
Master of Science in Bioinformatics, 2009

December 2017
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Dr. Andy Pereira
Dissertation Director

Dr. Vibha Srivastava
Committee Member

Dr. Ravi Barabote
Committee Member

Dr. Andrew Alverson
Committee Member

Abstract

Present day genomic technologies are evolving at an unprecedented rate, allowing interrogation of cellular activities with increasing breadth and depth. However, we know very little about how the genome functions and what the identified genes do. The lack of functional annotations of genes greatly limits the post-analytical interpretation of new high throughput genomic datasets. For plant biologists, the problem is much severe. Less than 50% of all the identified genes in the model plant *Arabidopsis thaliana*, and only about 20% of all genes in the crop model *Oryza sativa* have some aspects of their functions assigned. Therefore, there is an urgent need to develop innovative methods to predict and expand on the currently available functional annotations of plant genes. With open-access catching the 'pulse' of modern day molecular research, an integration of the copious amount of transcriptome datasets allows rapid prediction of gene functions in specific biological contexts, which provide added evidence over traditional homology-based functional inference. The main goal of this dissertation was to develop data analysis strategies and tools broadly applicable in systems biology research.

Two user friendly interactive web applications are presented: The Rice Regulatory Network (RRN) captures an abiotic-stress conditioned gene regulatory network designed to facilitate the identification of transcription factor targets during induction of various environmental stresses. The *Arabidopsis* Seed Active Network (SANe) is a transcriptional regulatory network that encapsulates various aspects of seed formation, including embryogenesis, endosperm development and seed-coat formation. Further, an edge-set enrichment analysis algorithm is proposed that uses network density as a parameter to estimate the gain or loss in correlation of pathways between two conditionally independent coexpression networks.

Acknowledgements

This dissertation would not be complete without thanking all the people who have directly and indirectly supported me in this exciting research endeavor. First and foremost, I would like to express my sincere appreciation to my advisor, Dr. Andy Pereira for supporting me throughout the course of this research. His immense experience and perspectives on solving problems of the future guided me to approach this research in the best possible way. His enthusiasm towards science, openness to new ideas and encouraging a collaborative environment are the traits I admire. Thank you, Dr. Andy, for taking a chance with me and letting me work with freedom and collaborate within and outside the lab. Your encouraging words immensely motivated me to do better than what I had started with, and your teachings will always guide my scientific pursuits.

I sincerely thank Dr. Arjun Krishnan at the Michigan State University for inspiring me with his work. His perspicuous concepts makes him one of the best contemporary scientist in the field of computational biology in my eyes. Every video call with Arjun has been enlightening and thought provoking. Thank you Arjun, for being open to sharing your ideas and offering me a position in your new lab. I look forward to being your student and wish you all the success with the goals you have set for yourself.

Special thanks to a very good friend and colleague, (soon-to-be) Dr. Ritika Mihani, without whom finding this great place to research would not have been possible. Ritu, I wish you all the success for your personal life and the professional journey you have embarked upon.

I would like to thank all the past and present members of the Pereira lab for creating an environment where my transition from human biology to plant biology was made easy. Thank you to the senior members of the lab for their intellectual inputs and collaborations, especially Dr.

Ramegowda Venkategowda for enhancing my knowledge about plant physiology and answering all those dull-witted questions I had on a number of occasions, Dr. Julie Thomas for her persistent support in data analysis, for critical review my manuscripts and thoughtful discussions on new perspectives and Dr. Supratim Basu for adding a new perspective to my research idea and sharing his deep background knowledge of rice molecular biology. I thank Dr. Subodh Srivastava for sharing his experiences as a senior Bioinformatician. I would also like to thank my fellow graduate student colleagues and soon-to-be doctorates, Anuj and Yehni for letting me work in their experimental fields and helping me understand drought. I thank Jawaher , Ipeleng and Mira for the discussions of their own projects and data that helped me understand plant molecular biology and physiology a little better. Last but not the least, I extremely thankful to Sara for always being there as a support system in the lab and making things easily accessible.

I thank all my dissertation committee members for their intellectual inputs on the projects and feedback during committee meetings. The candidacy exam and critical assessment of the proposal was a major learning point as a graduate student, and it motivated me to delve deeper into the biological side of bioinformatics. Thank you for always being open to my idiosyncrasies (e.g. accepting two abstracts for the candidacy instead of three) and showing light on many occasions when taking a decision seemed tough. I thank Dr. Andrew Alverson for the practical programming course I took with him helped me automate workflows and accelerate research.

I would like to thank the staff of Arkansas High Performance Computing Center for their computational support in analyzing the data used in this research. Special thanks to Dr. Pawel Wolinski for helping me set up the webserver.

I would like to thank all the teachers at the St. Mary’s school, Mt. Abu, for shaping my mind right from the beginning, and all the faculty members at the Sardar Patel University, especially Ms.

Pratibha Patil, Dr. Hetal Panchal and Dr. Utpal Dholakia for teaching me the basic skills of bioinformatics research.

I thank a dear friend, Ananya, from the bottom of my heart. Her moral support, patience and words of encouragement motivated me to a level that cannot be described in words. Thank you, Ananya, for always standing by my side.

This acknowledgement would not be complete without thanking my mother and my sister who believed in me more than I did in myself. Thank you mummy, for all the little success I have till now and all that to come in future is all because of you. You dedicated your life to struggle for our education, and I dedicate my life to give you all the happiness that you deserve. Thank you Palak, your unconditional love is what keeps me going.

And finally, thanks to that one person who made everything look easy and possible. Thank you for your measureless support and love from the Heavens. This one is for you, Papa!

Dedication

I dedicate this dissertation to my mother for her undying faith, patience and endless support in chasing my dreams.

Table of Contents

List of Submitted Papers

Chapter 3: SANe: The Seed Active Network for Mining Transcriptional Regulatory Programs of Seed Development
Chirag Gupta, Arjun Krishnan, Eva Collakova, Pawel Wolinski, Andy Pereira
(*Submitted to the bioRxiv server; doi: https://doi.org/10.1101/165894*)

Chapter 1: Introduction

The central goal of bio-molecular research is understanding how a genome functions and what cellular roles do the identified genes perform. With the advent of modern genomic technologies, the gap between the available DNA sequence and knowledge about the functions of these sequences is only widening. For example, using high-throughput sequencing technologies like RNA sequencing (RNA-seq), it is possible to quantify the regions of DNA that are transcribed, allowing one to measure the expression of genes and quantify their differences with varying experimental conditions. If the functional roles of genes that differentially express under a certain treatment as compared to control are known, identification of biological pathways/processes implicated in the treatment become apparent. However, even the most widely studied organisms do not have majority of the genes mapped to corresponding cellular functions, hampering and limiting the interpretation of a high-throughput experimental output. For plant biologists, the problem is much severe: even in the model plant *Arabidopsis thaliana*, only about 30% of all the identified genes have functional annotations that could be determined using the current state of art in molecular biology and genetics. Crops like rice, maize and wheat have less than 20% genes with functional annotations, including the gene functions that were predicted using computational approaches. In rice, genes whose functions were experimentally determined are less than 1%, and this number is still untraceable for other crops from information stored in public databases (Rhee and Mutwil, 2014).

With the tools of bioinformatics at the disposal of experiments in molecular biology, several computational and statistical methods are applied to genomic data that aid in functional assignments of genes in a predictive but high-throughput manner. For example, in homology based methods, sequence of a gene with unknown function is compared with similar sequences from

other organisms for which the function is known, since function conservation of homologs is an accepted theory (Whisstock and Lesk, 2003). Similarly, functional protein domains available in databases like PFam and DcGO are also sought in newly characterized protein sequences (Finn et al., 2010; Fang and Gough, 2013). As the structure is generally thought to be more conserved than the sequence, structural similarity between orthologous proteins also serves as a good indicator of function conservation (Sleator and Walsh, 2010). However, these analytical methods pose several limitations in understanding of gene function. Despite sharing a great degree of similarity in sequence, genes can evolve for very dissimilar functions (Gerlt and Babbitt, 2000). Likewise, proteins with similar functions can have very different structures (Omelchenko et al., 2010). Such inherent idiosyncrasies of gene function cannot be readily dealt with by comparing proteins or genes at the sequence level. Nevertheless, sequence conservation models provide an added level of information to support evidence from more dynamic models of gene function prediction.

The most reliable approach of function characterization follow sophisticated genetic screening protocols developed for several model organisms (Page and Grossniklaus, 2002; Kile and Hilton, 2005). Genetic screens are usually low-throughput, labor intensive and generally focusing on already established hypotheses surrounding a small number of genes or proteins known for involvement in a biological process. Moreover, it is an experimentally established fact that genes do not function in isolation, but groups of genes work together in an intricate network for a biological phenomenon/processes to occur. In such a functional group, disruption of a single gene might not always lead to an observable phenotype (White et al.; Bouche and Bouchez, 2001; Kok et al., 2015), due to genetic robustness acquired by dosage or genetic compensation (Tautz, 1992; El-Brolosy and Stainier, 2017). Further, given the amount of gene duplications a eukaryotic

genome holds, and the associated likelihood that a non-precise, random mutagenesis (or similar) strategy will disrupt a gene whose function is non-compensable, is very low.

The intricate network of genes is complex, multi-layered and changes with time, tissue and both endogenous and exogenous stimuli. The fundamental goal of systems biology is understanding the nature and dynamics of complex biological systems – how different components of the system work individually, and how they contribute to the functioning of the system as a whole. Network biology holds a key position in studies aimed at systems level understanding of an organism behavior, for example, in response to stress.

Network guided gene function prediction

A network – or graph in mathematical language – is composed of nodes and edges. In biological networks, a node can be any biological entity such as genes, proteins, a set of functionally coherent genes, diseases, traits etc., and edges are relationships or associations between these entities. In molecular datasets, nodes are most often genes or gene products, connected by edges if they interact/associate in a biological system. The edges can indicate a biophysical interaction, for example in protein interaction networks which are experimentally derived by yeast 2 hybrid assays or other similar *in vivo* large scale protein interaction assays (Schwikowski et al., 2000; Ding et al., 2009; Wu et al., 2010; Wei et al., 2014). The edges can also be associative, for example, genes exhibiting high degree of similarity in expression. A network of protein-DNA interactions can be modelled using high-throughput Chromatin immuno-Precipitation followed by sequencing (ChIP-seq), or medium throughput yeast one hybrid assays and immuno-precipitation assays to monitor which proteins bind to specific fragments of DNA, for example, the promoter regions of genes (Taylor-Teeples et al., 2015; Fuxman Bass et al., 2016).

The topology of biological networks, i.e. the organization of nodes and edges, has been observed to be very similar to that of other naturally occurring networks. For example, social networks, network of the World Wide Web or the network of genes in a cell, all have an inherent scale free topology (Barabasi and Albert, 1999). In a network with scale free topology, there are few nodes with a large number of connections and a large number of nodes with very few connections. Biologically, this type of topology provides efficiency in signal propagation (Klemm and Bornholdt, 2005; Peter and Davidson, 2017), as well as renders the network robust in handling random perturbations and maintaining cellular wellness (Hu et al., 2016). Hypothetically, genes that are prone to random mutations, or the selectively neutral iso-alleles (King and Jukes, 1969), are more likely to have their functions compensated (e.g. by gene duplications), thus limiting the mutation effect from propagating across a larger part of the network. Owing to higher rates of duplication events in plants (Panchy et al., 2016), redundancy is one of the reasons for mutants of several genes showing no discernible phenotypes (Barbaric et al., 2007). On the other hand, disruption of highly connected nodes will have a larger effect on the network behavior when perturbed. Hence, the 'hub' genes – genes with large number of connections in the network – are most sought after in network based gene prioritization in systems biology and translational research.

Network inference through integration of heterogeneous genomic datasets paints a more mechanistic picture of the working of the cell, and indicate the flow of biological information (Galperin and Koonin, 2000; Mostafavi and Morris, 2010; Davila-Velderrain et al., 2015). For example, heterogeneous networks (hetnets) developed from human knowledgebase represent different types of nodes (bio-molecular entities such as genes, tissues, disease etc.) and different

types of edges (carrying annotations about the specific biological relationship it represents) (Himmelstein and Baranzini, 2015).

In plants, especially in all the crops, we still lack enough heterogeneity in available molecular datasets, with expression data the most abundant type of data available for integration. Thus, the most prevalent type of gene networks in plants currently belong to the form of gene coexpression networks modelled using correlation of gene expression in a large sets of integrated microarrays (Ruan et al., 2010; Childs et al., 2011; Sato et al., 2012; Takehisa et al., 2015) or RNA-seq (Iancu et al., 2012; van Dam et al., 2015). Several online platforms for gene network inference using 'associalogs' (conserved functional linkages) are available for plants (Katari et al., 2010; Gu et al., 2011; De Bodt et al., 2012; Franceschini et al., 2013). However, there is still a lack in availability of 'context-specific' gene networks to generate credible hypothesis, considering versatility of homolog gene function in plants.

Community detection and imputing functions from gene networks

Biological networks are complex and possess remarkable amount of structure. Community detection, or clustering is the most powerful technique in network inference. Clustering offers a method to break down large networks into manageable groups. The groups are clusters which represent genes that connect with each other more than they connect to genes in other clusters, and can be thought of as part lists or modules with common functions that can be employed by the cell on need basis by selectively turning a module on or off according to the developmental phase or environmental cue. The biological function of a cluster can be determined by statistically testing its overlap with gene lists of known function (Castillo-Davis and Hartl, 2003). Hence, this cluster-wide propagation of functional annotations, or 'guilt-by association', automatically provides

putative annotations to genes with no previous knowledge about their biological functions, but are in neighborhood of genes with known functions.

Several graph clustering algorithms have been devised for detecting such gene communities in a network. The SPICi algorithm clusters a network based on density threshold (Jiang and Singh, 2010), assuming that the gene connections within a biological relevant community are denser as compared to connections between genes from different unrelated communities. The same concept is explored by the Louvain algorithm that optimizes modularity and reveals the best possible grouping in the data by a greedy search process (Vincent, 2008). The Topological Overlap Matrix detects modules based on the adjacency matrix (correlation matrix) and the topological features (weighted degrees) of each gene (Dong and Horvath, 2007). The Markov Cluster Algorithm (MCL) is an unsupervised algorithm based on stochastic flows in graphs, and controls the granularity of clusters using an inflation parameter *I* (van Dongen and Abreu-Goodger, 2012). Regardless of which clustering algorithm to use, different values of clustering parameters used by the algorithm should be first evaluated using a 'gold-standard' for their ability to cluster genes that are already known to work as a group. Such groupings can be obtained from functional annotation catalogs from different ontologies and pathway databases. Hence, an extensive data-driven approach to evaluate clustering could reveal the best parameter for the underlying dataset (Krishnan et al., 2017).

Gene function prediction from integrated transcriptomes of similar biological context

It is well documented that genes that are transcriptionally coordinated tend to be functionally related (Mutwil, 2011). This transcriptional coordination between genes can be estimated by integrating a diverse set of gene expression datasets to model a biological network. Such a network

is correlational, and connects gene-pairs if they have a high degree of correlation in their expression. Aptly termed ‘coexpression networks’ (CN), functional predictions from such networks can be traced back to the beginning of last decade and were based on assumptions that a statistically significant coexpressed gene-pair might possibly have shared regulatory inputs, and thus can be functionally related (D’haeseleer et al., 2000). ‘Guilt-by association’ using microarrays became a popular theory (Quackenbush, 2003; Wolfe et al., 2005) and gained wide popularity due to availability of gene expression data in large scale from the public domains like the GEO database (Barrett et al., 2007). Figure 1.1 depicts the typical steps involved in a standard coexpression network inference in systems biology. With certain limitations to the uses and interpretation of CN (Gillis and Pavlidis, 2012), numerous studies have used CN for gene function prediction in various organisms such as humans (Lee et al., 2004; Elo et al., 2007; Prieto et al., 2008), mouse (Menashe et al., 2013; Liu and Ye, 2014), bacteria (Jiang et al., 2016), yeast (van Noort et al., 2004) and a variety of other model systems (Stuart et al., 2003).

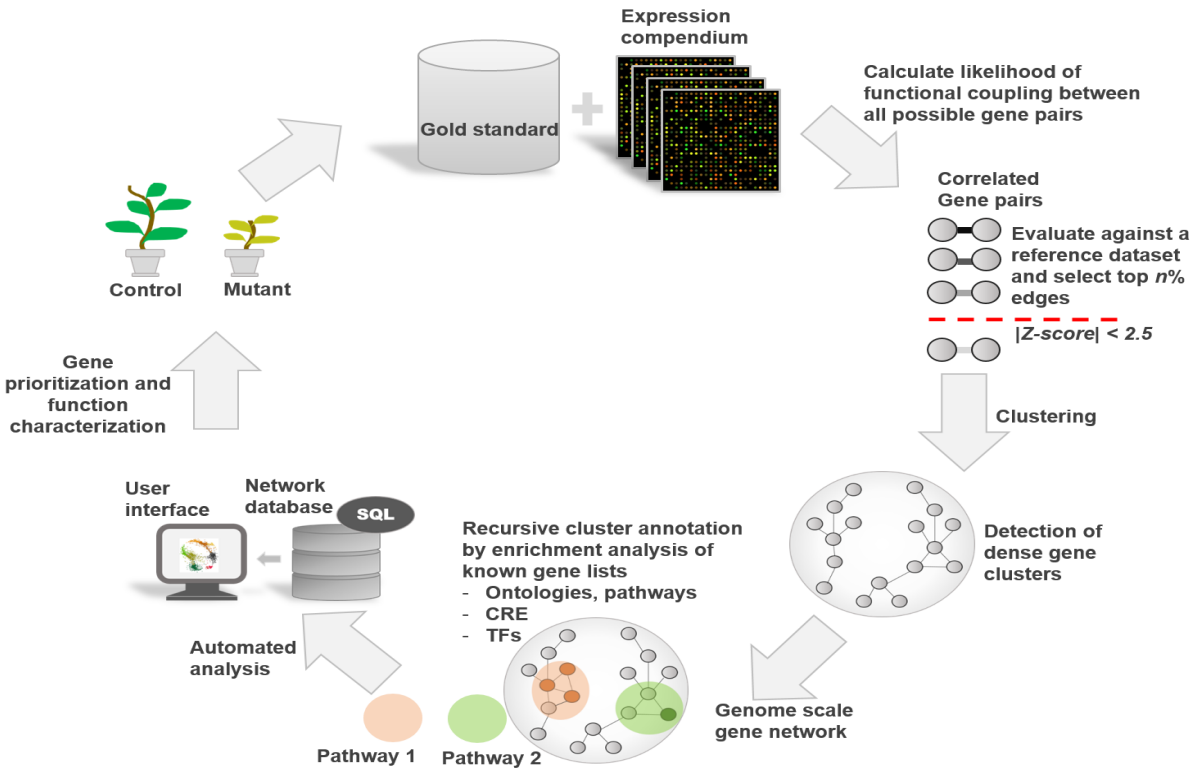

Figure 1.1: Coexpression network mining workflow. A coexpression network is derived using a large compendium of expression datasets. The datasets can be sampled from public repositories like the NCBI GEO, and should be chosen to represent a unifying biological context (e.g. response to abiotic-stress, specific tissues). Correlations in expression profiles of all possible gene pairs across all the samples is then calculated, typically using Pearson's Correlation or Mutual Information based statistical models. Statistically significant edges (gene pairs) are selected and evaluated using a 'gold standard' or reference dataset to test the robustness of clustering or test the accuracy of predictions in regulatory network inference. The resulting network is a list of gene pairs with edge scores indicating the magnitude of their covariance or likelihood of their functional relatedness, depending upon the correlation measure used. The network is then scanned for occurrence of gene clusters that are dense, meaning genes that form a community and 'interact' with each other more than with genes outside their respective cluster. Such clusters represent genes with common cellular goals, and are most often coregulated by the same sets of TFs. The biological factor causing such functional coherence between genes in a cluster is determined by statistically calculating its overlap with gene lists of known functions (e.g. a list containing all the genes involved in photosynthesis, cell cycle etc.). Most often, not all genes within a predicted cluster have a functional annotation, but if the cluster is significantly associated with a biological process, the functions of these unknown genes can be imputed, thus adding additional genes to the original sets of functional gene lists. The data is presented to the scientific community under creative commons license via the web for free, and a community driven approach is taken to use the data to prioritize genes for experimental validations. Newly validated gene functions are then added to the gold standard for further refining the computational predictions in future experiments.

Information content in coexpression networks

The edges in a CN network are usually quantified using statistical models that generally fall in two categories: correlation and Mutual Information (MI). In correlation models, Pearson Correlation (PC) is the most popular statistical measure for associating gene pairs. However, one has to bear in mind that PC, by design, does not prove causality between variables, and in terms of gene pairs, it can only state the degree to which two genes associate with each other, calculated from their covariance in a bivariate Gaussian distribution. Moreover, PC can only model relationships that are linear, which is not always the case with genes. Some genes can relate with each other only in certain tissue/organs, stress conditions, developmental phases etc. Thus, the datasets chosen for integration with an aim of coexpression have to be close to a common biological context to make the interpretations more robust, and with sample size large enough to satisfy the underlying assumptions of this model.

Another popular method of connecting genes on the basis of expression patterns is using Mutual Information (MI). MI is an information theoretic procedure useful in detecting genes that have a non-linear pattern of coexpression (e.g. genes correlated in a subset of samples and not in the rest), and states causality to some extent (Steuer et al., 2002). MI, however, is computationally expensive for the analysis of larger genomes, considering the fact that it requires a very large sample size for estimation as it involves data discretization (usually $\sqrt{N}$ bins, where N is the sample size), and permutations to score for significance. While both PC and MI yield similar results, the range of scores is very different. PC scores range from -1 to 1, indicating negative and positive correlations, respectively. Absolute PC values close to 1 indicate stronger associations. On the other hand, MI scores can have only positive values ranging from 0 to infinity, with larger scores

indicating stronger likelihood of the two genes ‘interacting’. Comparatively, both PC and MI yield very similar results, and a high PC cannot yield a low MI and vice versa.

Thus, CN can only be used to estimate associations between two genes in terms of their expression. Unlike protein-protein interaction networks, CN network do not explicitly state that the protein products of two highly correlated genes physically interact *in vivo*. The strength of CN analysis lies in its framework; besides expanding the available ontologies for functional enrichment analysis (Gupta et al., 2017; Krishnan et al., 2017), coexpression scores state the degree of functional coupling between genes for a weighted analysis framework that opens avenues for integration of CN with heterogeneous genomic datasets.

There is no good consensus among scientists on which statistical models agree with the 'true network', based on a ‘gold standard’ created from an existing knowledgebase. However, the 'best' or most cited method might not work optimally for the dataset in hand, especially in cases when the software has never been tested for a broad range of organisms and datasets. A pioneering work at the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), has shown that the best current solution is to use results of an ensemble of methods, and aggregate results in an unbiased manner (Marbach et al., 2012). This scheme has worked well in terms of accuracy of analysis shown in the evaluation plots for network analysis in subsequent studies.

Chapter 2 focuses on using an ensemble of methods to predict an abiotic-stress specific transcriptional regulatory network of the crop model rice (*Oryza sativa*). A large sample size of ~595 samples from ~29 different datasets was used for predicting the targets of TFs in the rice genome. Following the methods described by Marbach et al., the results from four distinct reverse-engineering solutions were aggregated into a ‘consensus network’. In addition to that, using

benchmarking schemes, it was shown that removal of the worst performer increased the accuracy of the final aggregate predictions. Additionally, the same method was using to identify the best and worst performers for analyzing RNA-seq datasets. The CuffDiff, edgeR, limma and DESeq2 algorithms were evaluated for their ability to detect a set of *bonafide* drought related genes. Further, of the seven new TFs predicted to regulate heat stress response in rice, the biological role of one was experimentally confirmed. The interactive webserver for this project is developed for flexible network mining of rice transcriptome datasets.

Chapter 3 is a seed-specific transcriptional regulatory network referred to as SANe, for Seed Active Networks. The network was created using transcriptomic datasets generated from different stages of seed development in Arabidopsis, using a modified version of the best performer (CLR) algorithm from evaluations in rice. The algorithm correctly predicted TFs that are already known for their functional roles in development of different seed parts, such as the embryo, endosperm and seed coat regions, and further suggested new TFs for experimentation *in vivo*. Mutants of a few TFs of interest from the prediction set were acquired and tested for accumulation of seed storage compounds like starch, oil and proteins – the factors that contribute most to the economic potential of plants. The network is currently stored as a MySQL database with an interactive user interface accessible at https://plantstress-pereira.uark.edu/SANe/. The platform is integrated with several tools that will enable seed biologists to analyze their own new datasets and generate new testable hypothesis regarding genes involved in seed development

Chapter 4 proposes a method for 'differential networking', in which a control network is compared to a drought-specific network in rice. The algorithm first calculates the coherence of known functional gene sets in both the networks, and estimates gene-pairs that have significantly rewired functional interactions, leading to network-density based pathway-level fold change

estimation rather than fold change of individual genes. The results suggests that the algorithm reveals several pathways that are truly associated with drought, but could not be retrieved by traditional methods, and further suggested drought candidates that are differentially coexpressed under drought but not differentially expressed.

References

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. Nat Genet **25**

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science **286:** 509-512

Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BMJ, Kappos L, Gene MSAC, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Human Molecular Genetics **18:** 2078-2090

Barbaric I, Miller G, Dear TN (2007) Appearances can be deceiving: phenotypes of knockout mice. Briefings in Functional Genomics **6:** 91-103

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucl Acids Res **35**

Bouche N, Bouchez D (2001) Arabidopsis gene knockout: phenotypes wanted. Curr Opin Plant Biol **4:** 111-117

Castillo-Davis CI, Hartl DL (2003) GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics **19:** 891-892

Chan EK, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. PLoS Biol **9:** e1001125

Childs KL, Davidson RM, Buell CR (2011) Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. PLoS ONE **6:** e22196

Davila-Velderrain J, Martinez-Garcia JC, Alvarez-Buylla ER (2015) Descriptive vs. mechanistic network models in plant development in the post-genomic era. Methods Mol Biol **1284:** 455-479

De Bodt S, Hollunder J, Nelissen H, Meulemeester N, Inze D (2012) CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. New Phytol **195:** 707-720

Ding X, Richter T, Chen M, Fujii H, Seo YS, Xie M, Zheng X, Kanrar S, Stevenson RA, Dardick C, Li Y, Jiang H, Zhang Y, Yu F, Bartley LE, Chern M, Bart R, Chen X, Zhu L, Farmerie WG, Gribskov M, Zhu J-K, Fromm ME, Ronald PC, Song W-Y (2009) A Rice Kinase-Protein Interaction Map. Plant Physiology **149:** 1478-1492

Dong J, Horvath S (2007) Understanding network concepts in modules. BMC Systems Biology **1:** 24-24

D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics **16:** 707-726

El-Brolosy MA, Stainier DYR (2017) Genetic compensation: A phenomenon in search of mechanisms. PLOS Genetics **13:** e1006780

Elo LL, Järvenpää H, Orešič M, Lahesmaa R, Aittokallio T (2007) Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics **23:** 2096-2103

Fang H, Gough J (2013) DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res **41:** D536-544

Ficklin SP, Feltus FA (2013) A systems-genetics approach and data mining tool to assist in the discovery of genes underlying complex traits in Oryza sativa. PLoS One **8:** e68551

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A (2010) The Pfam protein families database. Nucleic Acids Research **38:** D211-D222

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res **41**

Fuxman Bass JI, Pons C, Kozlowski L, Reece-Hoyes JS, Shrestha S, Holdorf AD, Mori A, Myers CL, Walhout AJM (2016) A gene-centered C. elegans protein–DNA interaction network provides a framework for functional predictions. Molecular Systems Biology **12:** 884

Galperin MY, Koonin EV (2000) Who's your neighbor? New computational approaches for functional genomics. Nat Biotech **18:** 609-613

Gerlt JA, Babbitt PC (2000) Can sequence determine function? Genome Biology **1:** reviews0005.0001-reviews0005.0010

Gillis J, Pavlidis P (2012) "Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks. PLOS Computational Biology **8:** e1002444

Gu H, Zhu P, Jiao Y, Meng Y, Chen M (2011) PRIN: a predicted rice interactome network. BMC Bioinformatics **12:** 161

Gupta C, Krishnan A, Collakova E, Wolinski P, Pereira A (2017) SANe: The Seed Active Network For Mining Transcriptional Regulatory Programs of Seed Development. bioRxiv

Himmelstein DS, Baranzini SE (2015) Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. PLOS Computational Biology **11:** e1004259

Hu JX, Thomas CE, Brunak S (2016) Network biology concepts in complex disease comorbidities. Nat Rev Genet **17:** 615-629

Iancu OD, Kawane S, Bottomly D, Searles R, Hitzemann R, McWeeney S (2012) Utilizing RNA-Seq data for de novo coexpression network inference. Bioinformatics **28:** 1592-1597

Jiang J, Sun X, Wu W, Li L, Wu H, Zhang L, Yu G, Li Y (2016) Construction and application of a co-expression network in Mycobacterium tuberculosis. **6:** 28422

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

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res **32**

Katari MS, Nowicki SD, Aceituno FF, Nero D, Kelfer J, Thompson LP, Cabello JM, Davidson RS, Goldberg AP, Shasha DE, Coruzzi GM, Gutierrez RA (2010) VirtualPlant: a software platform to support systems biology research. Plant Physiol **152:** 500-515

Kile BT, Hilton DJ (2005) The art and design of genetic screens: mouse. Nat Rev Genet **6:** 557-567

King JL, Jukes TH (1969) Non-Darwinian Evolution. Science **164:** 788

Klemm K, Bornholdt S (2005) Topology of biological networks and reliability of information processing. Proceedings of the National Academy of Sciences of the United States of America **102:** 18414-18419

Kok FO, Shin M, Ni CW, Gupta A, Grosse AS, van Impel A, Kirchmaier BC, Peterson-Maduro J, Kourkoulis G, Male I, DeSantis DF, Sheppard-Tindell S, Ebarasi L, Betsholtz C, Schulte-Merker S, Wolfe SA, Lawson ND (2015) Reverse genetic screening reveals poor correlation between morpholino-induced and mutant phenotypes in zebrafish. Dev Cell **32:** 97-108

Korber N, Bus A, Li J, Higgins J, Bancroft I, Higgins EE, Parkin IA, Salazar-Colqui B, Snowdon RJ, Stich B (2015) Seedling development traits in Brassica napus examined by gene expression analysis and association mapping. BMC Plant Biol **15:** 136

Krishnan A, Gupta C, Ambavaram MMR, Pereira A (2017) RECoN: Rice Environment Coexpression Network for Systems Level Analysis of Abiotic-Stress Response. bioRxiv

Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res **14**

Liu W, Ye H (2014) Co-expression network analysis identifies transcriptional modules in the mouse liver. Molecular Genetics and Genomics **289:** 847-853

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. Nat Meth **9:** 796-804

Menashe I, Grange P, Larsen EC, Banerjee-Basu S, Mitra PP (2013) Co-expression Profiling of Autism Genes in the Mouse Brain. PLOS Computational Biology **9:** e1003128

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

Mutwil M (2011) Integrative transcriptomic approaches to analyzing plant co-expression networks. PhD Thesis: Integrative Ansätze zur Analyse von Koexpressionsnetzwerken in Pflanzen. University Potsdam.

Omelchenko MV, Galperin MY, Wolf YI, Koonin EV (2010) Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. Biol Direct **5:** 31

Page DR, Grossniklaus U (2002) The art and design of genetic screens: Arabidopsis thaliana. Nat Rev Genet **3:** 124-136

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

Peter IS, Davidson EH (2017) Assessing regulatory information in developmental gene regulatory networks. Proc Natl Acad Sci U S A **114:** 5862-5869

Prieto C, Risueño A, Fontanillo C, De Las Rivas J (2008) Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles. PLOS ONE **3:** e3911

Quackenbush J (2003) Microarrays--Guilt by Association. Science **302:** 240

Rhee SY, Mutwil M (2014) Towards revealing the functions of all genes in plants. Trends in Plant Science **19:** 212-221

Ruan J, Dean AK, Zhang W (2010) A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Systems Biology **4:** 8

Sato Y, Namiki N, Takehisa H, Kamatsuki K, Minami H, Ikawa H, Ohyanagi H, Sugimoto K, Itoh J-I, Antonio BA, Nagamura Y (2012) RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Research

Schwikowski B, Uetz P, Fields S (2000) A network of protein-protein interactions in yeast. Nat Biotech **18:** 1257-1261

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**

Sleator RD, Walsh P (2010) An overview of in silico protein function prediction. Arch Microbiol **192:** 151-155

Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics **18:** S231-S240

Stuart JM, Segal E, Koller D, Kim SK (2003) A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science **302:** 249

Takehisa H, Sato Y, Antonio B, Nagamura Y (2015) Coexpression Network Analysis of Macronutrient Deficiency Response Genes in Rice. Rice **8:** 1-7

Tautz D (1992) Redundancies, development and the flow of information. Bioessays **14:** 263-266

Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R, Handakumbura PP, Xiong G, Wang C, Corwin J, Tsoukalas A, Zhang L, Ware D, Pauly M, Kliebenstein DJ, Dehesh K, Tagkopoulos I, Breton G, Pruneda-Paz JL, Ahnert SE, Kay SA, Hazen SP, Brady SM (2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature **517:** 571-575

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. Plant J **37:** 914-939

van Dongen S, Abreu-Goodger C (2012) Using MCL to extract clusters from networks. Methods Mol Biol **804:** 281-295

van Noort V, Snel B, Huynen MA (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Reports **5:** 280-284

van Dam S, Craig T, de Magalhães JP (2015) GeneFriends: a human RNA-seq-based gene and transcript co-expression database. Nucleic Acids Research **43:** D1124-D1132

Vincent DBaJ-LGaRLaEL (2008) Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008:** P10008

Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet **11:** 843-854

Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, Cartinhour S, Stein LD, McCouch SR (2002) Gramene, a tool for grass genomics. Plant Physiol **130**

Wei S, Hu W, Deng X, Zhang Y, Liu X, Zhao X, Luo Q, Jin Z, Li Y, Zhou S, Sun T, Wang L, Yang G, He G (2014) A rice calcium-dependent protein kinase OsCPK9 positively regulates drought stress tolerance and spikelet fertility. BMC Plant Biol **14:** 133

Weng L, Macciardi F, Subramanian A, Guffanti G, Potkin SG, Yu Z, Xie X (2011) SNP-based pathway enrichment analysis for genome-wide association studies. BMC Bioinformatics **12:** 99

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys **36:** 307-340

White JK, Gerdin A-K, Karp NA, Ryder E, Buljan M, Bussell JN, Salisbury J, Clare S, Ingham NJ, Podrini C, Houghton R, Estabel J, Bottomley JR, Melvin DG, Sunter D, Adams NC, Baker L, Barnes C, Beveridge R, Cambridge E, Carragher D, Chana P, Clarke K, Hooks Y, Igosheva N, Ismail O, Jackson H, Kane L, Lacey R, Lafont David T, Lucas M, Maguire S, McGill K, McIntyre RE, Messager S, Mottram L, Mulderrig L, Pearson S, Protheroe HJ, Roberson L-A, Salsbury G, Sanderson M, Sanger D, Shannon C, Thompson PC, Tuck E, Vancollie VE, Brackenbury L, Bushell W, Cook R, Dalvi P, Gleeson D, Habib B, Hardy M, Liakath-Ali K, Miklejewska E, Price S, Sethi D, Trenchard E, von Schiller D, Vyas S, West AP, Woodward J, Wynn E, Evans A, Gannon D, Griffiths M, Holroyd S, Iyer V, Kipp C, Lewis M, Li W, Oakley D, Richardson D, Smedley D, Agu C, Bryant J, Delaney L, Gueorguieva NI, Tharagonnet H, Townsend AJ, Biggs D, Brown E, Collinson A, Dumeau C-E, Grau E, Harrison S, Harrison J, Ingle Catherine E, Kundi H, Madich A, Mayhew D, Metcalf T, Newman S, Pass J, Pearson L, Reynolds H, Sinclair C, Wardle-Jones H,

Woods M, Alexander L, Brown T, Flack F, Frost C, Griggs N, Hrnciarova S, Kirton A, McDermott J, Rogerson C, White G, Zielezinski P, DiTommaso T, Edwards A, Heath E, Mahajan MA, Yalcin B, Tannahill D, Logan DW, MacArthur DG, Flint J, Mahajan VB, Tsang SH, Smyth I, Watt FM, Skarnes WC, Dougan G, Adams DJ, Ramirez-Solis R, Bradley A, Steel KP (2013) Genome-wide Generation and Systematic Phenotyping of Knockout Mice Reveals New Roles for Many Genes. Cell **154:** 452-464

Wolfe CJ, Kohane IS, Butte AJ (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics **6:** 227

Wu G, Feng X, Stein L (2010) A human functional protein interaction network and its application to cancer data analysis. Genome Biology **11:** R53

Chapter 2: An abiotic-stress conditioned gene regulatory network of rice predicted using an ensemble of reverse engineering algorithms

Abstract

Drought and other environmental stressors trigger several metabolic and physiological changes in plants, and a part of this change is reflected in the transcriptome. It is known that a synergistic coordination of the gene-regulatory machinery senses, modifies and controls the desired metabolic state of the cell. The exploitation of genes that regulate stress response has been hampered by low genetic and phenotypic evidence. Hence, there is a critical need for innovative and complementary approaches to identify the components of gene regulatory networks that manifest during abiotic stress. In this study, a new web-application referred to as the Rice Regulatory Network (RRN) is presented. RRN reports regulatory interactions predicted using an ensemble of four reverse-engineering algorithms and covers ~62% of all the identified genes models in rice. RRN was evaluated using the existing knowledgebase of rice, in which it consistently ranked known DNA binding sites of TFs and other experimentally validated TF-gene interactions towards the top of all the predicted ranks. RRN can be searched by aligning the differential expression values in new transcriptomes to a set of 'coregulated' clusters, or by searching a list of biologically coherent genes to prioritize TFs for further experimentation.

Introduction

Plants, like all multicellular organisms, respond to environmental stimuli by complex mechanisms that regulate the transcription of thousands of genes. These changes in gene expression are controlled by regulatory proteins at multiple levels. A transcribed mRNA molecule can be selectively processed or degraded, or post-translation modifications can render a protein molecule active or inactive in accordance with the metabolic needs of the cell. These complex gene regulatory mechanisms which involve interactions and cooperation of several kinases, phosphatases and transcription factors (TFs) serve as an efficient mechanism for signal transduction for flexible metabolic states, while maintaining cellular hemostasis. Identifying such regulatory genes is the key to understanding the fundamentals of stress response in plants that can aid in molecular engineering of crops with economic importance. Rice, as an example of model for cereals, is an agriculturally important crop feeding almost half of the world population. However, the fraction of stress response genes known in rice is very small compared to validated genes in the model plant Arabidopsis.

TFs are an important class of regulatory proteins and have a strong evidential support as regulators of stress and development (Charu et al.; Joshi et al., 2016). TFs regulate the expression of other genes, regarded as their targets, by binding to their promoters regions at specific sites known as *cis* regulatory elements (CRE). At the transcriptional level, some information about TF mediated gene regulation can be deduced from genome-wide expression profiles obtained using microarrays or RNA-sequencing (RNA-seq). Since TFs are themselves transcriptionally regulated, studying their expression patterns and the patterns of correlated genes often reveals a suite of their *in vivo* targets. An integrative analyses of large-scale gene expression datasets from multiple experimental conditions, typically in the form of coexpression networks, uncovers the dynamics

of gene activity that often remains hidden in an individual experiment. This technique has now become a useful and powerful approach in plant systems biology for prediction and validation of novel gene functions (Lee et al., 2009; Ficklin et al., 2010; Childs et al., 2011; Sato et al., 2012). Many online platforms are available that allow one to interrogate the biological roles of interesting genes using coexpression as the basis. Since functionally related genes have similar dynamics of expression, guilt-by association helps in associating functions to uncharacterized genes based on the known functions of their neighboring genes in an underlying network.

Although coexpression networks provide a useful way to group functionally similar genes, edges (connected genes) in coexpression network merely represent the degree of similarity in their expression profiles. It has been observed that cascades of transcriptional interactions tend to correlate expression of many downstream genes, many of which may not necessarily interact physically. However, in a wet-lab setting, direct regulatory interactions are of most interest, especially those involving TFs capable of activating or repressing a functional program. Moreover, coexpression between genes can be non-linearly associated with time and growth stages. The regulatory potential of a gene can be over or underestimated if the phenomenon of indirect regulation is not taken into account (Gordân et al., 2009). Hence, gene regulatory networks (GRN) emerged as a class of biological networks where edges are solely between TFs and putative targets and indicate a possible causal relationship. Many statistical solutions for inference of regulatory networks from expression data have been proposed, which take into account known TF genes to infer regulatory interactions that are direct (Yu et al., 2004; Margolin et al., 2006; Faith et al., 2007; Joshi et al., 2009).

This study is an attempt to assemble an abiotic-stress conditioned GRN of rice. Expression data derived under the context of ‘abiotic-stresses’ were downloaded from public repositories and

integrated to calculate coexpression between genes using Mutual Information (MI) as the statistical measure of correlation. MI was used to maximize the estimation of causal dependency between genes. MI provides a generalized measure of correlation between genes and is more sensitive to non-linear relationships, unlike the Pearson's Correlation (Steuer et al., 2002). Further, an ensemble of four published reverse-engineering methods was created, and each one supplied with MI scores to estimate the statistical likelihood of TFs directly interacting with potential target genes, i.e. filtering non-direct interactions. The resulting data served two main purposes; identification of potential ranked targets of TFs, and the discovery of TFs that potentially coregulate the common sets of targets genes. The latter was quantified on gene pair basis and used to perform a weighted clustering analysis to identify functionally coherent gene groups. The identified clusters were annotated with functional and regulatory information that greatly expanded upon the available functional annotations of rice genes. The Rice Regulatory Network (RRN) is developed to integrate the network data with analysis tools that can be used through the web, and provides an easy to use access to experimental biologists for functional analysis of their own datasets and prioritize genes for further experimentation.

In rice and many cereals, the developing inflorescence with flowers that will bear seed after fertilization, and the flag leaf supporting the inflorescence with nutrition for development, are essential factors determining grain yield under drought and other stresses. To address the biological functions involved as an example, rice inflorescence and flag leaf tissues treated to drought that reveal drought responsive genes by RNA-seq analysis, are used to illustrate the network properties and applications of the network (Fig 2.1).

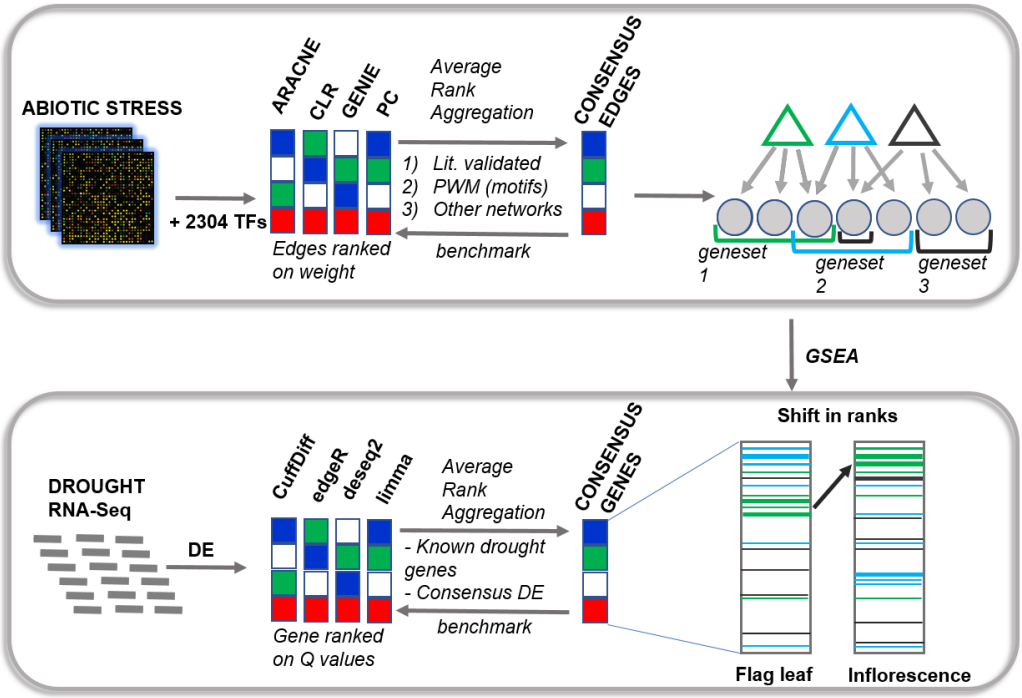

Figure 2.1: Workflow for mining transcriptional regulatory programs from an abiotic-stress expression compendium of rice. Top: Building and benchmarking a consensus network using an ensemble of four reverse-engineering algorithms to predict the targets of transcription factor genes. The toy-bars represent edges predicted by each of the four methods used in the ensemble, where each colored box represents the rank of a unique edge. ARACNE: Algorithm for Reconstruction of Accurate Cellular Networks, CLR: Context Likelihood of Relatedness, GENIE: GEne Network Inference Engine, PC: Pearson's Correlation. Bottom: Evaluation of different software for the analysis of rice Flag leaf and Inflorescence drought response RNA-seq dataset, and integration with the predicted network model using a modified gene set analysis algorithm. The toy-bars represent genes predicted by each of the four methods used in the ensemble, where each colored box represent the rank of a unique gene.

Results

Combining predictions from different algorithms to assemble the regulatory network

A set of 29 publicly available Affymetrix based gene expression datasets of rice were downloaded from the Gene Expression Omnibus (GEO) database (Barrett et al., 2007). These datasets comprised of a total of 266 samples (595 Affymetrix GeneChips) from experiments performed under a specific biological context, 'response to environmental stress', which included experiments from drought, salt, heat and hormone stresses. Along with a list of 2304 genes known to function as TFs identified from three different databases (Yilmaz et al., 2009; Jung et al., 2010; Priya and Jain, 2013), expression values of 35,151 rice genes were supplied to four complementary reverse-engineering algorithms. Three of these methods are products of the DREAM5 challenge (Dialogue for Reverse Engineering Assessment Methods) and have been shown to accurately predict different parts of the underlying regulatory network (Margolin et al., 2006; Faith et al., 2007; Huynh-Thu et al., 2010). Since these methods used MI to quantify coexpression, a linear correlation based method using Fisher's Z transformed Pearson's Correlation (PC) scores was also added to the ensemble (Huttenhower et al., 2006) for comparison of the prediction outcomes. From each of the four methods, the top 500,000 predicted edges (TF-target interactions) were selected and ranked based on the confidence weights. As expected, the union of top 500,000 edges of all the four methods showed very little overlap (< 1%) and resulted in a total of ~1.5 million edges (Fig. 2.2). This overlap was slightly better than that observed in the Arabidopsis consensus network (Vermeirssen et al., 2014).

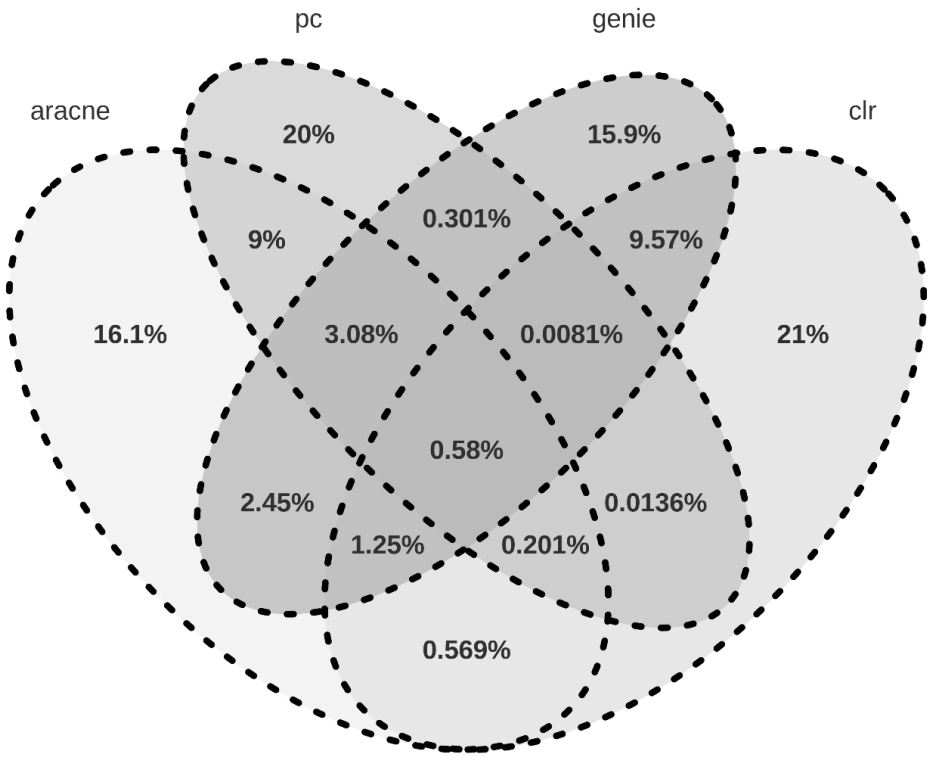

Figure 2.2: Venn diagram illustrating the percentage of ~1.5 million edges predicted by each of the four network prediction tasks and their overlaps. ARACNE: Algorithm for Reconstruction of Accurate Cellular Networks, CLR: Context Likelihood of Relatedness, GENIE: GEne Network Inference Engine, PC: Pearson's Correlation.

Integrating predictions from different methods have been shown to have a better predictive power as compared to using an individual method alone, as shown in bacteria and yeast (Marbach et al., 2012). Therefore, an ensemble solution was computed based on the average of ranks for each edge, and the top 500,000 edges of the aggregate were selected (see "Methods"). The average rank aggregation method kept those edges on the top that were ranked higher by a larger fraction of individual methods in the ensemble. Hence, edges predicted by only a few methods were automatically suppressed towards a lower rank keeping the false positives to a minimum (Marbach et al., 2012). The average rank aggregation has also been recently used in the validated Arabidopsis cell wall regulatory network (Taylor-Teeples et al., 2015) and has been shown to perform better than other methods of aggregation, like the union and mean-reciprocal aggregation (Vermeirssen et al., 2014).

Evaluating the biological significance of each algorithm and their ensemble

A *Gold Standard (GS)* for evaluation of predictions is hard to create in plants other than the well-studied model, Arabidopsis. Typically, the most stringent *GS* is created by using a set of functional annotations with experimental validation, which being sparse in rice, greatly limits the evaluation test on coverage of novel predictions. Since this study was aimed at predicting direct targets of TFs, the best possible *GS* for evaluation would be a set of experimentally validated TF-target interactions in the literature. Extensive manual and automated literature mining of such interactions (at the time the study was conducted, roughly till 2014) identified only 308 interactions between 114 TFs and 194 target genes, a number too low for evaluation of a large-scale genomic study like the one presented here.

This limitation of an effective *GS* in rice was in part overcome by building three independent non-expression-data based reference networks (RN) for evaluation of each prediction algorithm and their ensemble solutions. The first RN was a set of 308 experimentally validated edges mentioned above. The second RN was a combination of high-confidence edges in the RicenetV2 (Lee et al., 2015), the Predicted Rice Interactome Network (PRIN) database (Gu et al., 2011) and the String database of protein interactions (Szklarczyk et al., 2015), resulting in a total of 213,523 edges between 2007 TFs and 20,518 target genes. For the third RN, the position weight matrices of rice TFs listed in the CIS-BP database (Weirauch et al., 2014) were obtained, and these TFs were linked to genes which harbored the corresponding motifs in their promoter regions (see "Methods"). Hence, a total of 219,460 edges between 587 TFs and 21,080 genes formed the third RN. Since a large fraction of the edges in all the three RNs represent interactions that were reported independent of gene expression, considering these as 'known' and evaluating overlaps with the edges predicted solely from expression data in this study served as benchmarks to estimate the performance of the algorithms.

For benchmarking each prediction method and their ensemble solution against the RNs, an F-score was calculated from the values of precision (number of correctly predicted edges) and recall (number of known edges that were predicted) (see "Methods"). As expected, a very low F-score was observed for all the evaluations (Table 2.1). Within this narrow range of F-scores, the CLR algorithm outperformed all other methods in the second and third RNs, including the ensemble solution, and covered a larger fraction of known edges in the first RN. This observation of CLR as the best performer in an ensemble is in agreement with the Arabidopsis study (Vermeirssen et al., 2014), and hence establishes itself as one of the most suitable method, at least for plant gene expression datasets. On the other hand, PC based method performed the worst in

any given RN, validating that TF and target genes are infact non-linearly associated. Moreover, when interactions from PC method were removed from the ensemble and the aggregation re-computed (Avg-sans-PC), the F score of the ensemble solution increased significantly in all the RNs, with a 2-fold increase for edges that were experimentally determined in the first RN (Table 2.1).

Table 2.1: Benchmarking the predictions. Test in performance of the top 500,000 predictions from the four individual algorithms and their ensemble solution, on correctly predicting known regulatory edges. An F score was computed as F= 2*(Precision * Recall)/ Precision + Recall, where Precision = TP/ TP + FP, and Recall = TP/TP + FN. Higher F scores indicate a higher level of accuracy in predictions. Here, TP, TN, FP and FN stands for True Positives, True Negatives, False Positives and False Negatives, respectively.

Reference	# Predicted	TP	Precision	Recall	F
Experimentally validated (308 edges)					
ARACNE	5461	2	0.00649	0.00036	0.001
CLR	34379	20	0.06	0.0005	0.001
Genie	23299	14	0.045	0.0006	0.001
PC	23309	2	0.006	8.00E-05	0
Avg. Rank	22721	17	0.05	0.0007	0.001
Avg-sans-PC	20927	16	0.0526	0.0007	0.002
Predicted-string, PRIN, ricenet2 (213523 edges)					
ARACNE	298568	3082	0.0144	0.0103	0.012
CLR	478824	16188	0.0758	0.0338	0.047
Genie	398162	11249	0.0526	0.0282	0.037
PC	395463	1404	0.0065	0.0035	0.005
Avg. Rank	395437	11324	0.053	0.0286	0.037
Avg-sans-PC	390208	12467	0.0583	0.0319	0.041
Motif-association net (219460 edges)					
ARACNE	38101	1249	0.0056	0.0327	0.01
CLR	130271	2273	0.0103	0.0174	0.013
Genie	81832	1603	0.0073	0.0195	0.011
PC	97663	1646	0.0075	0.0168	0.01
Avg. Rank	89592	1980	0.009	0.0221	0.013
Avg-sans-PC	85189	1944	0.008	0.0228	0.013
Union network (432495 edges)					
ARACNE	305465	4321	0.014145647	0.009990867	0.012
CLR	483924	18414	0.03805143	0.042576215	0.04
Genie	402465	12825	0.031866125	0.029653522	0.031
PC	401011	3045	0.007593308	0.007040544	0.007
Avg. Rank	400107	13257	0.033133637	0.030652377	0.032
Avg-sans-PC	395128	14360	0.036342654	0.033202696	0.035

Further, how much confidence a method placed in predicting an edge was evaluated on the basis of how well the edges were ranked. For this test, a union of 432,495 edges was taken by combining the edges in all the three RNs. Intersection of this union and the edges in each of the four prediction methods as well as their ensemble solution was evaluated for the rank-positions. It was observed that the median-rank of the edges in the Avg-sans-PC aggregation was the lowest, followed by the ensemble aggregation that still had interactions derived from the PC based method. The third quartile of ranks covered by Avg-sans-PC was just slightly over the top 50% of all the edges, with the median lying within the first three deciles of all the rankings (Fig. 2.3). This indicated that the Avg-sans-PC aggregate of predictions were robust and the most reliable aggregation to make biological interpretations. This aggregate, comprising of ~500,000 edges between 2282 TFs and 33,876 target genes was chosen as the 'consensus network' and sorted by the ranks of TF-gene interaction predicted, with smaller ranks indicating high confidence edges. The consensus network is referred to as the 'Rice Regulatory Network' (RRN) and is made available on an online interactive browser available under open access at https://plantstress-pereira.uark.edu/RRN/ .

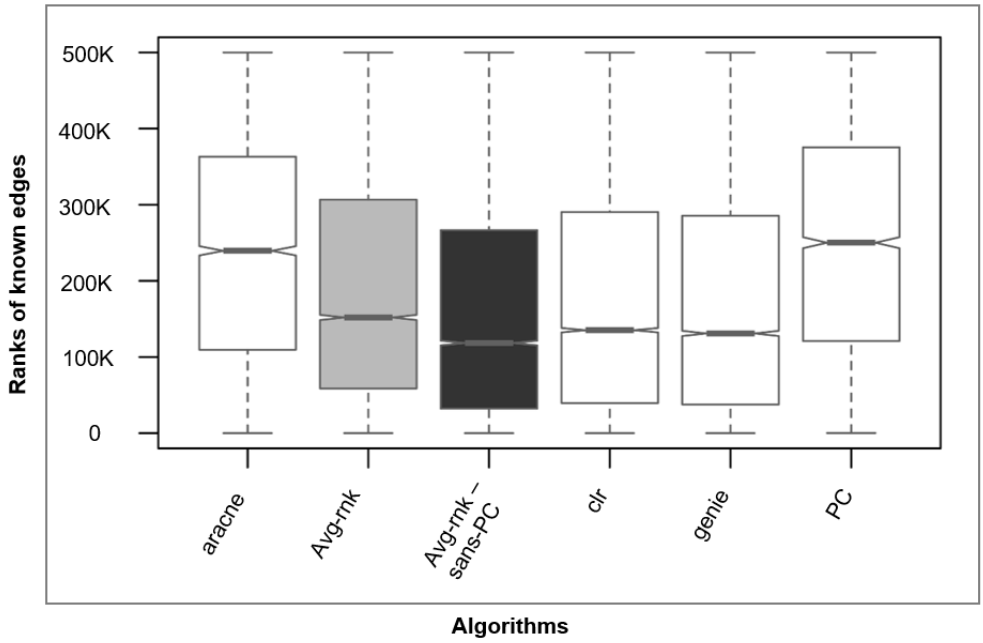

Figure 2.3: Evaluation of the network robustness. A set of known edges from three distinct reference networks was taken, and their ranks evaluated in each of the four prediction tasks and their ensemble (X axis). The overall ranks were plotted as boxplots and the center of the box corresponds to the median (2^{nd} quartile; Q_2) of the distribution of ranks (Y axis). 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), and indicate a 95% confidence interval for the difference in two means. The boxes of two ensemble solutions are shaded as grey for the average rank aggregation method and black for the average rank – sans – Pearson's correlation (PC) based method. Lower medians indicate more robust predictions.

Clustering the coregulatory network to detect functionally coherent modules

Clustering is a method of identifying 'communities' or clusters of functionally coherent genes (D'Haeseleer, 2005). In terms of biological networks, clustering detects genes that connect with each other more densely than other genes outside their respective clusters (Jiang and Singh, 2010). Known functional annotations of genes within a given cluster can then be propagated cluster-wide, automatically annotating uncharacterized genes within the cluster assuming a 'guilt-by association'.

To cluster the consensus network derived above (the RRN), which is essentially a regulatory network, coregulation between genes was first detected by computing overlaps in their predicted regulators (TFs). This method of clustering coregulated genes served an alternative to clustering based on coexpression. Since coexpressed genes can also be coregulated and shared targets of a TF can function in a similar cellular pathway (Yu et al., 2003), this method is expected to reveal genes that are under the control of same set(s) of TFs, adding an additional layer of regulatory information unlike coexpression based clustering methods. Coregulation between genes was computed using the Jaccard's Index (JI) of similarity, and a threshold of JI 0.01 was set to connect 27,004 genes with each other. The Markov Cluster Algorithm (MCL) was then employed to find clusters of genes that are more densely connected to each other than with genes outside their clusters (van Dongen and Abreu-Goodger, 2012), using JI scores as edge-weights. MCL is based on stochastic flows in graphs assuming the natural property that random walks within a dense cluster will likely result in visiting most of the nodes within the cluster as compared to clusters that are not so dense. MCL requires a single inflation parameter *I* that controls the granularity of clusters. Instead of choosing an arbitrary value if *I*, clusters obtained at a range of *I* values were evaluated for biological process (BP) categories from the Gene Ontology (GO)

database (see “Methods”).

While the total number of clusters with significantly enriched BP categories varied with different *I* values, the number of distinct BP enriched in the clusters gradually decreased with increasing values of *I* (Fig. 2.4 A). The overall accuracy, measured as geometric mean of Positive Predictive Value (PPV: the ability of a cluster to detect a BP) and Sensitivity (Sn: how well genes belonging to the same BP are grouped in the same cluster) decreased with increasing values of *I*. Cluster separation, indicating the bidirectional correspondence between a BP category and a given cluster, increased till *I* 2.25 and remained almost constant thereafter (Fig. 2.4 B). Apart from *I* 2.25 indicating a slight better separation than *I* 2.0, all other parameters indicated that the biological information of the clusters is best preserved at *I* 2.0. Hence, an *I* value of 2.0 was set as the inflation parameter to obtain 881 clusters of coregulated genes. Coregulated clusters thus obtained were annotated with GO BP categories, pathways from the KEGG catalog of rice and known *cis*-regulatory elements (CRE) of plants (see “Methods”). A TF was labeled as a regulator of a cluster if at least 50% of its predicted targets lie within the cluster (Vermeirssen et al., 2014).

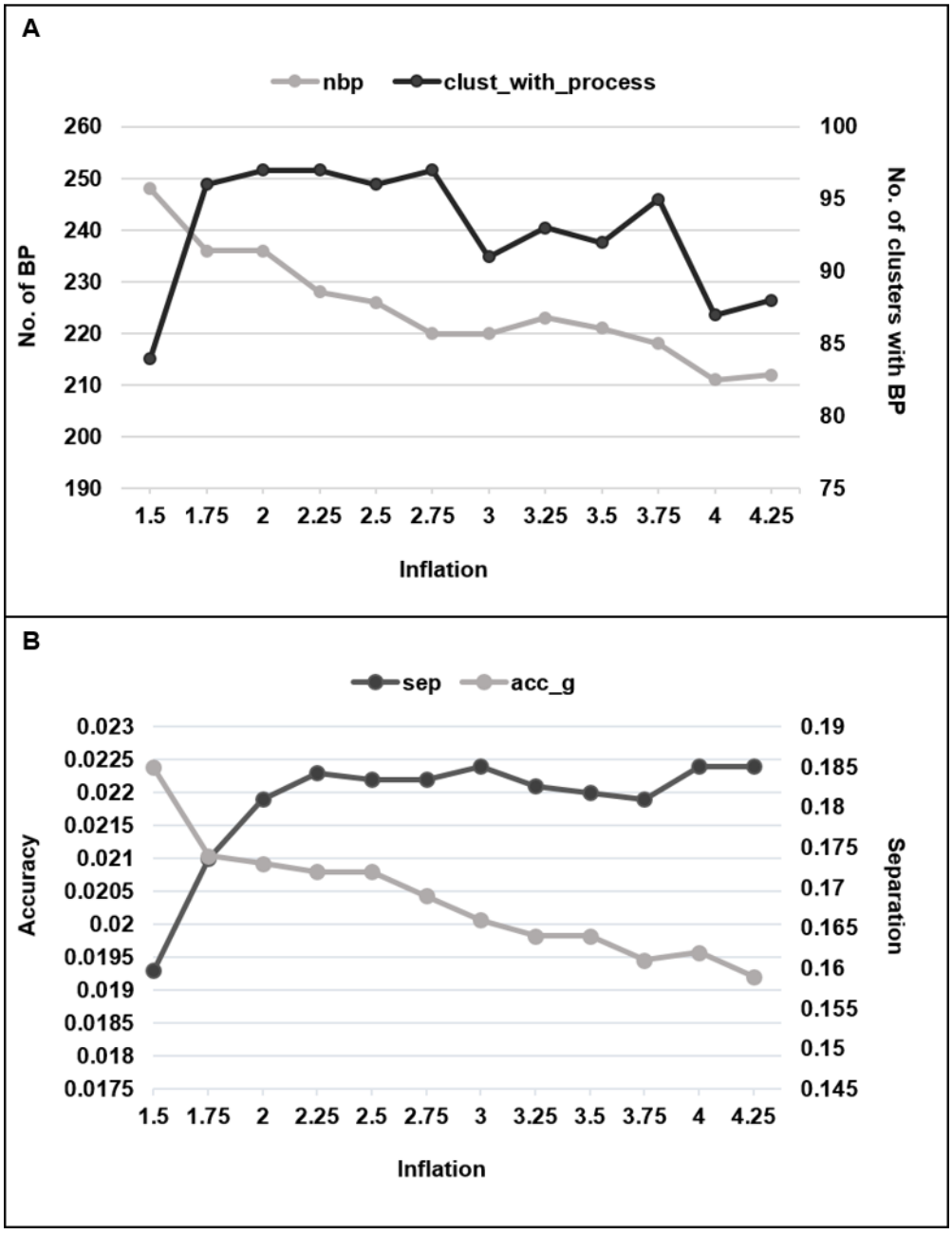

Figure 2.4: Evaluation of inflation parameter *I* of the MCL algorithm for clustering. A range of inflation parameters (Y axis) required by the MCL algorithm were tested for: A) Total number of biological processes (BP) detected as significant, and the total number of clusters with at least one significantly enriched BP. The *p*-value of enrichment was computed under a cumulative hypergeometric test, and an FDR corrected *p*-value of < 0.05 was set as a significance threshold. B) Evaluation of different *I* values for the overall accuracy (calculated as the geometric mean of the positive predictive value and sensitivity in detecting functional complexes on the left Y axis) and separation (indicating how well the members of a BP are grouped into one cluster on the right Y axis).

Cluster 128 is a heat stress module dominated by the HSF regulon with evidence of post-transcriptional regulation

Cluster 128 (cl_128) comprises of a total of 37 genes significantly enriched with genes in GO BP category “protein folding”, “response to abiotic stimulus” and “response to heat”. Significantly overrepresented pathways in cl_128 were related to “Protein processing in endoplasmic reticulum” and “Spliceosome”, according to rice KEGG pathway annotations. 5 of the 7 predicted regulators of this cluster belong to the class of Heat Shock Factors: HSFA2A (LOC_Os03g53340), HAP2A (LOC_Os08g09690), HSFA2C (LOC_Os10g28340), HSFB2A (LOC_Os04g48030) and HSFB2C (LOC_Os09g35790), indicating an HSF dominated regulon triggered under heat stress. Interestingly, ~80% of all the predicted targets of one of the regulator designated as splicing factor U2af (LOC_Os09g31482) belong to this cluster, explaining the observation of KEGG term spliceosome. Another interesting regulator of this module is ethylene-responsive transcriptional coactivator/endothelial differentiation-related factor 1 (LOC_Os06g39240) with two known alternatively spliced isoforms. The Arabidopsis homolog of this gene is the Multiprotein bridging factor-1 (AT3G24500) that acts as a transcriptional coactivator.

Geneset enrichment of RNA-seq data with coregulated gene clusters

The typical protocol for analysis of an expression dataset leads to finding pathways or biological processes that are most highly associated with the observed response or phenotype in the experiment. This analysis is sometimes limited when the available functional annotations are sparse and incomplete, as in the case with rice. The main aim of clustering genes was to expand these available functional annotations, so that the identified clusters can be used in a geneset

enrichment analysis (GSEA) framework in contrast to directly using genesets from GO or KEGG annotations. In the case of RRN, the enrichment analysis will also be helpful in finding the regulators of the clusters that are most significantly associated with the input expression data, a piece of information that is usually hidden in GO based enrichment analysis. The RRN webserver is integrated with an enrichment analysis tool (Kim and Volsky, 2005) to detect clusters most highly associated with an uploaded transcriptome. An example of such an analysis is shown below.

Querying RRN with RNA-seq data of rice Flag leaf and Inflorescence tissue in response to drought

In rice and many cereals, the developing inflorescence with flowers that will bear seed after fertilization and the flag leaf supporting the inflorescence with nutrition for development (Li et al., 1998), are essential factors determining grain yield under stress. RNA-seq analysis of Flag Leaf (FL) and Inflorescence (INF) tissue in response to drought was conducted to identify associated clusters and their regulators in the drought sensitive reproductive tissue. Reads obtained from the paired-end sequencing of two biological replicates for each of the control and drought treatments for both the tissue samples were aligned to the MSUv7 rice reference transcriptome using the Tophat v 2.0.12 based Bowtie aligner (Trapnell et al., 2009). The concordance between two replicates of each sample was evaluated as the read coverage in bins of genomic regions (Ramirez et al., 2014). The Pearson's correlation coefficient value between replicates of the FL tissue was 0.94 and of the INF sample was 0.98. Since the ultimate goal was to correspond RNA-seq reads with network-data, reads were counted at gene level instead of isoform level (Liao et al., 2014).

To evaluate the level of induction or repression in genes with drought treatment as a factor,

the four most widely used statistical models, Cuffdiff (Trapnell et al., 2013), edgeR (McCarthy et al., 2012), deseq2 (Love et al., 2014) and limma (Ritchie et al., 2015) were evaluated for performance. Genes were first ranked on the basis of their *p/q* values obtained from each method. The performance of each method was then tested by detecting the ranks of 5419 *bona fide* drought genes (DG). This DG list was created by taking the intersection of differentially expressed genes in six independent drought microarray experiments available in rice (GSE21651, GSE24048, GSE25176, GSE26280 and GSE81253). Since these experiments covered a wide spectrum of different drought treatments and rice genotypes, an intersection of genes that differentially expressed in these experiments are guaranteed to be drought responsive. Under this framework, it was observed that all the methods were generally in agreement with each other for both the samples. However, the edgeR algorithm was found to perform better as the median ranks of DG was the lowest (Fig. 2.5). EdgeR even performed better than the ensemble solution computed using the 'average ranks aggregation' method, as done for the consensus regulatory network. Hence, FC values estimated using edgeR were uploaded into the RRN webserver to find the enrichment of coregulated clusters and associated functional and regulatory information in both the samples.

A total of 56 clusters were found significantly enriched in at least one of the two samples (Fig 2.6). Of these clusters, 35 clusters had functional annotations that expand a wide variety of biological processes from the GO ontology database and pathways from the KEGG pathways (Fig 2.7). A few of these clusters are described below.

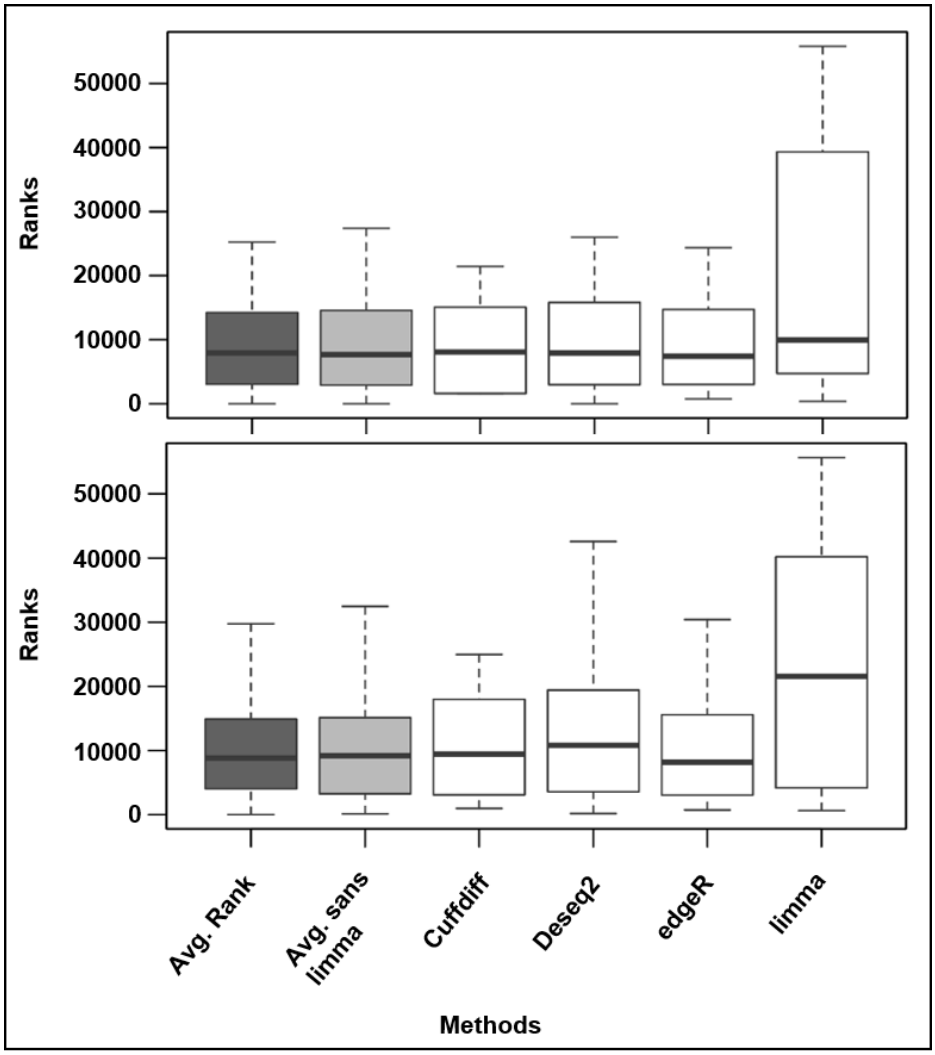

Figure 2.5: Evaluation of four different RNA-seq analysis methods. Four different published software for analysis of differential expression in RNA-seq data (X-axis) were evaluated on the basis of how well they rank a set of ~5000 *bona fide* Drought Genes (DG; see the main text on how these genes were identified) on the Y-axis. All the assayed genes were ranked on the basis of their FDR corrected *p*-values resulting from each method. The distribution of overall ranks of DG were plotted as boxplots, leaving out outlier values above the whiskers for clarity. The center of the box corresponds to the median (2nd quartile; Q_2) of the distribution and the extremes of the box correspond to the 1st (Q_1) and 3rd (Q_3) quartiles. The whiskers denote $Q_2 \pm 1.5*IQR$, where IQR is the interquartile range (Q_3-Q_1). Two ensemble solutions were also computed by taking the average of ranks across all the methods (dark grey box) and leaving out the worst performer from the ensemble (light grey box). Lower medians values indicate better performance in detecting DE.

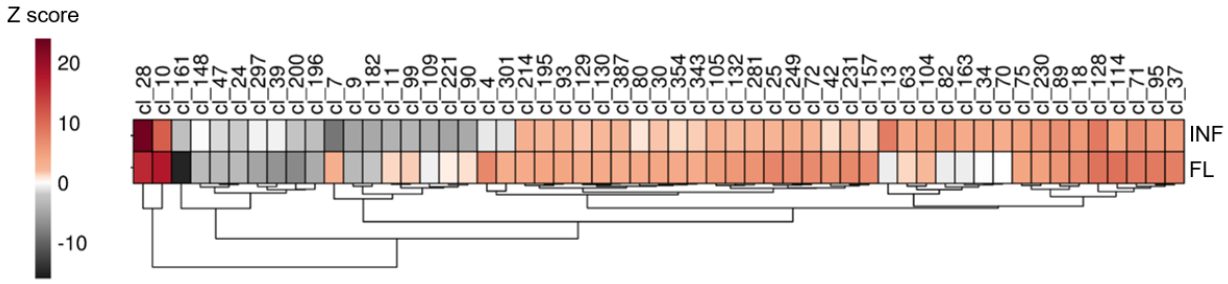

Figure 2.6: Coregulated clusters enriched in the drought transcriptomes of Flag leaf (FL) and inflorescence (INF) tissues. Log of fold change values upon exposure to drought computed using edgeR were used in the Parametric Analysis of Geneset Enrichment (PAGE) framework on the RRN webserver to detect clusters that are most significantly enriched in both the transcriptomes. The resulting table was downloaded and a heatmap of Z scores plotted in R. The heatmap shows two-tailed enrichment Z scores with red and grey gradient showing positive and negative enrichment, respectively.

Cluster ID	GO BP	KEGG Pathway
cl_10	carboxylic acid catabolism	PLANT-PATHOGEN INTERACTION
cl_104	polysaccharide metabolism	NA
cl_11	lipid localization	PENTOSE AND GLUCURONATE INTERCONVERSIONS
cl_128	response to heat	PROTEIN PROCESSING IN ENDOPLASMIC RETICULUM
cl_129	membrane lipid metabolism	NA
cl_13	chromatin modification	SPLICEOSOME
cl_132	cellular nitrogen compound metabolism	NA
cl_148	mRNA metabolism	NA
cl_161	protein folding	CIRCADIAN RHYTHM-PLANT
cl_163	purine nucleotide metabolism	NA
cl_18	protein transport	MRNA SURVEILLANCE PATHWAY
cl_182	cell cycle	NA
cl_196	cellular nitrogen compound biosynthesis	NA
cl_221	secondary metabolism	NA
cl_24	translational initiation	SPLICEOSOME
cl_25	response to drug	NA
cl_28	response to water stimulus	NA
cl_30	regulation of protein metabolism	UBIQUITIN MEDIATED PROTEOLYSIS
cl_34	protein catabolism	PROTEIN PROCESSING IN ENDOPLASMIC RETICULUM
cl_39	RNA modification	PURINE METABOLISM
cl_4	response to oxidative stress	PHENYLPROPANOID BIOSYNTHESIS
cl_42	response to light stimulus	SULFUR METABOLISM
cl_47	programmed cell death	NA
cl_63	protein catabolism	NA
cl_7	aromatic compound catabolism	PHENYLPROPANOID BIOSYNTHESIS
cl_72	metal ion transport	GALACTOSE METABOLISM
cl_75	vesicle-mediated transport	NA
cl_80	protein transport	NA
cl_82	ubiquitin-dependent protein catabolism	NA
cl_9	programmed cell death	HOMOLOGOUS RECOMBINATION
cl_90	generation of precursor metabolites and energy	NA
cl_93	homeostasis	NA
cl_99	oligopeptide transport	NA
cl_114	NA	FATTY ACID METABOLISM

Figure 2.7: Pathways and Biological Processes perturbed under reproductive stage drought. The first column shows the IDs of the clusters that were detected as significantly enriched in the FL and INF samples. The second and the third columns show Gene Ontology Biological Process terms (GO BP) and the KEGG pathway terms significantly enriched within these clusters, respectively. Only terms with the lowest *p*-values are shown. NA indicates that no term was detected significant under an FDR corrected *p*-value threshold of < 0.05.

Cluster 28 is a water stress module regulated by ABA

Cluster 28 (cl_28) comprises of 220 genes, most of which are up-regulated in both the tissues exposed to drought (Fig. 2.6). The GO BP terms overrepresented in the cluster are "response to water stimulus" and "response to abiotic stimulus", among others. The Abscisic acid responsive element identified in Arabidopsis ABREATRD22 (Iwasaki et al., 1995) and a calmodulin-binding GCGCBOXAT involved in stress signal transduction (Yang and Poovaiah, 2002) are significantly enriched in the promoters of genes in this cluster. 12 TFs were predicted as the regulators of this cluster: OsHSFA7, OsTZF8, HSFC2a, OsHOX24, OsERF95, OsTZF1, OsNAC88, OsABL1, OsHSF16, OsMYB7, LOC_Os12g06010 and LOC_Os11g06130. From these, OsHSFA7, OsTZF1 and OsABL1 are already known for their involvement in drought as suggested in the literature (Yang et al., 2011; Jan et al., 2013).

Cluster 7 is downregulated specifically in the inflorescence tissue

Cluster 7 is comprised of 133 genes and is enriched with genes annotated to "aromatic compound metabolism", "glucan metabolism" and "cell wall" related processes in the GO categories, and "Phenylpropanoid biosynthesis" and "Starch and Sucrose metabolism" pathways from the KEGG pathways. Among the predicted regulators of this cluster, targets of OsMYB58/63, the secondary cell wall related NAC29, MYB86 and BTB8 had the highest overlap. The association of MYB family TFs with these pathways during reproductive development has been shown previously (Wilson and Zhang, 2009), which also suggests that this cluster may function in the development of reproductive structures by altering the phenylpropanoid pathway, justifying its downregulation in the INF tissue.

Development of the RRN web application

The RRN and the definitions of coregulated clusters along with their functional and regulatory annotations are stored in a MySQL database. The database can be accessed at https://plantstress-pereira.uark.edu/RRN/ (Fig. 2.8). The current version of RRN incorporates two important features: the first feature allows users to enter a list of locus ID of interest and search coregulated clusters that are significantly represented in the list. The statistical significance of enrichment is calculated using a hypergeometric test (Castillo-Davis and Hartl, 2003) and clusters that stand a FDR corrected *p*-value > 0.05 are displayed to the user. The second feature allows one to upload a transcriptome to identify coregulated clusters that are significantly enriched with the phenotype. The significance of enrichment is computed using the Parametric Analysis of GeneSet Enrichment model (Kim and Volsky, 2005) as an integrated tool re-written in perl. Since the model uses parametric statistics to calculate *p*-values, phenotype values of all the genes assayed should be uploaded. The phenotype values can be either fold change of expression in response to a treatment for a two-sided enrichment test, or it can be *p/q* values of differential expression (DE) for a one sided enrichment test.

The resulting clusters from these analyses are linked to their gene table along with all the functional and regulatory annotations. Genes within each of the resulting clusters are displayed as a graph using Cytoscape web (Lopes et al., 2010). Gene and edge attributes in the graph are set to highlight information encoded in the expression file uploaded by the user. For example, nodes are shaped according to their molecular function, i.e. triangle to represent a TF, diamond to represent a kinase and ellipse to represent other genes (Fig. 2.9).

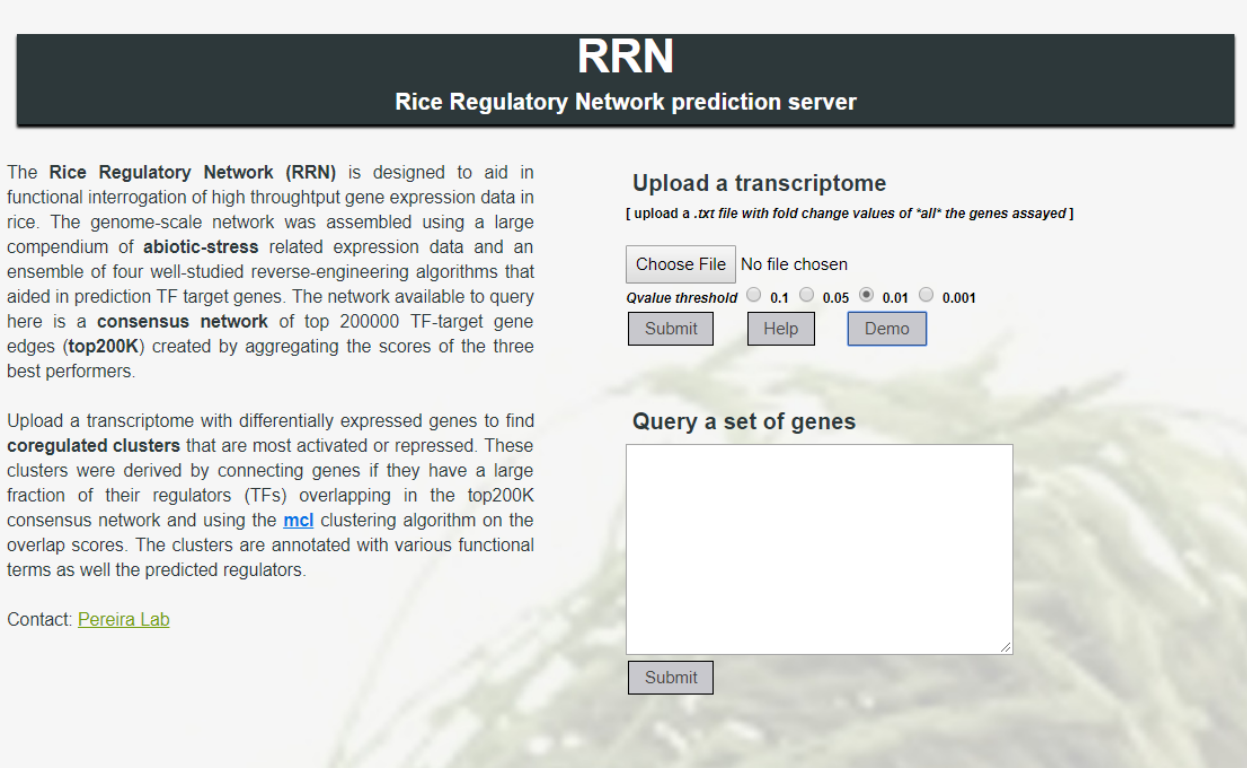

Figure 2.8: Screenshot of the RRN webserver.

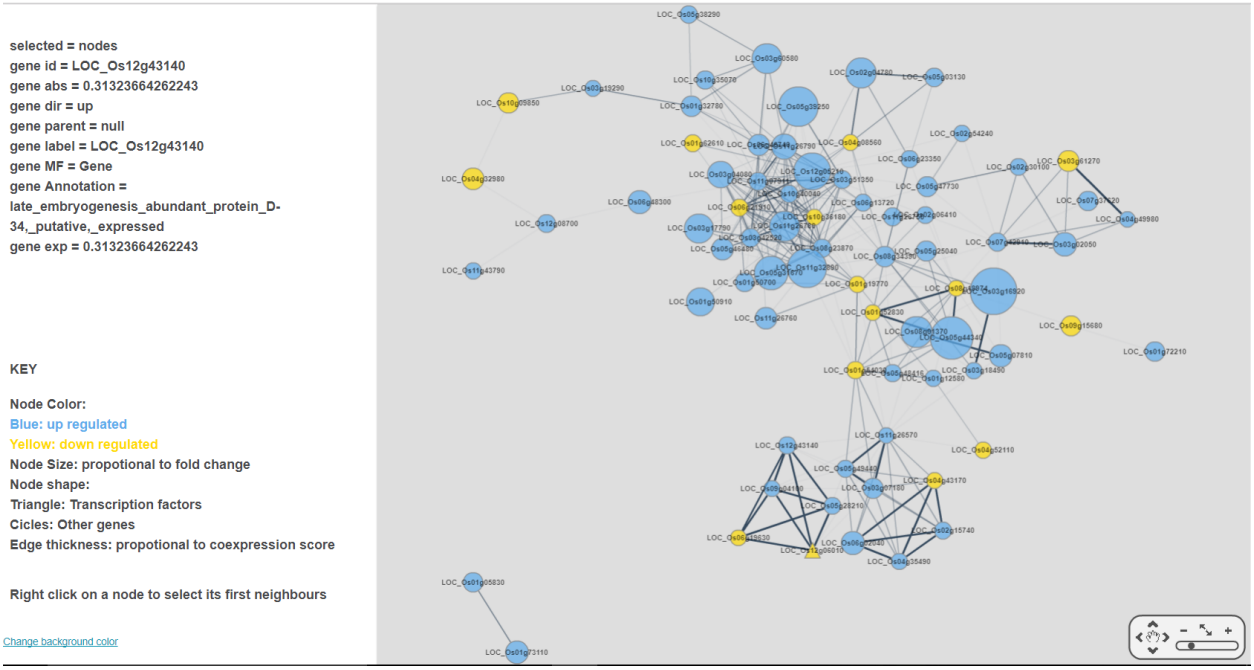

Figure 2.9: Screenshot of the cluster graph displayed on the RRN webserver.

Discussion

This study was the first attempt to reverse-engineer a genome-wide regulatory network of rice. Four complementary statistical algorithms were used to predict interactions between TFs and target genes using publicly available expression data conditioned on abiotic stresses. A suitable balance between precision and recall was achieved implicitly by choosing a 'belief based gold-standard' comprising of literature derived experimentally confirmed interactions, a set of known TF-DNA interactions, and high-confidence interactions predicted in other published rice networks, to benchmark the predicted regulatory network. Evaluations showed that an ensemble of these algorithms is more robust in predictions and provides a complete picture of the underlying regulatory landscape. The average rank aggregation method used to create the ensemble increased the accuracy of predictions overcoming the assumptions and biases introduced with use of a single method. Moreover, it was observed that removing the worst performer increases the accuracy of the ensemble solution even more, providing direct evidence against using linear models like the Pearson's Correlations as a measure to infer TF targets from expression data.

Furthermore, it was observed that clustering genes based on their coregulatory patterns reveals an additional layer of regulatory information that is usually hidden in coexpression based clustering. The identified clusters of coregulated genes were annotated with functional and regulatory information which stands more informative in gene set enrichment analysis of additional RNA-seq or microarray data, in addition to providing information about TFs that likely regulate associated biological processes and pathways. This utility of the network was shown by using the coregulated clusters as gene sets in enrichment analysis of drought induced FL and INF transcriptomes. The pipeline revealed several biological processes and pathways that are usually observed to be perturbed under drought, in addition to revealing several TFs with known stress

responses. Exploration of a few of such functional clusters leads to hypothesis generation and prioritize genes for experimental validation. For example, Cluster 128 revealed five interconnected heat shock factors that regulate heat stress response, which might also be under post-transcriptionally regulated as indicated by the prediction of the splicing factor U2af as one of the regulators.

It was observed that the average rank aggregation method cannot be generalized to analysis of any molecular datatype, as suggested by evaluations of an ensemble of RNA-seq data analysis methods on two stress-induced transcriptomes of rice. A logical reasoning surrounding this observation is based on the biological question itself and the underlying assumptions of the statistical models used. The statistical tests used for differential expression analysis simply state whether the change in the observed levels of gene expression in two samples is significant or not. While in network analysis, a large number of biological assumptions can be made but cannot be simultaneously integrated into a single statistical framework, hence taking the top confidence predictions from different methods into a consensus is more robust to inclusion of false positives.

Although using the consensus network scheme has been used before in plants (Vermeirssen et al., 2014; Taylor-Teeples et al., 2015) and other smaller organisms (Marbach et al., 2012), the study presented here has several distinguishing factors. There is not a single database available for analysis of additional newer expression data, as the existing network browsers are limited to gene/set query (Sato et al., 2012; Lee et al., 2015). The online interactive webserver of the RRN presented here is one of its kind and will be helpful in analysis of RNA-seq datasets for retrieving information about perturbed biological processes, pathways and most dominant TFs as regulators of the observed phenotype. Additionally, targets of a TF of interest can also be identified and validated using medium or high throughput protein-DNA binding *in vivo* assays (Basu et al., 2014)

for reconstruction of regulatory networks.

Methods

The abiotic-stress expression compendium

A set of 29 microarray datasets comprising of 595 samples were downloaded from the GEO database. Raw .CEL files were background corrected, normalized and summarized using the Robust Multiarray Average algorithm in R (Irizarry et al., 2003). A custom rice Chip Definition Format (CDF) file was used to assign Affymetrix probes to individual genes in the MSU V7 annotation, covering a total of 35,151 unique gene models. Replicates were averaged and individual expression matrices were combined to create an integrated expression matrix with 35,151 genes in rows and 266 samples in columns, with each cell in the matrix representing the log of intrinsic expression value of the corresponding gene in the corresponding sample column.

Ensemble solution for target-gene prediction.

A list of 2304 rice genes identified as TFs was curated from online databases. This list along with the expression matrix was supplied as input to ARACNE (Margolin et al., 2006), GENIE (Huynh-Thu et al., 2010) and CLR (Faith et al., 2007) algorithms to predict direct interactions between the TFs and target genes. The source code of ARACNE and GENIE software were downloaded from the published project pages, while the CLR version implemented in the R package minet (Meyer et al., 2008) was used. For the PC based model, correlation scores between all gene-pairs were mapped to Z scores and edges with at least one node as TF from the list were identified. For each of the four algorithms and the ensemble, edges were ranked based on the confidence scores and

the top 500,000 edges were selected. Next, the average of ranks for each edge was calculated. Ranks of edges that were not predicted by an algorithm was set to the total number of rankings +1 (~500,001) as the lowest rank. A custom R script was written to rank-organize the data and aggregate the results into a consensus.

Derivation of reference networks and benchmarking

For the second RN, the RicenetV2, PRIN and String interactions were downloaded. From each of these independently inferred networks, transcriptional edges (edges with at least one node from the list of TFs used in this study) were identified. A union of all such edges was then computed. For the third RN (motif-association net), the PWM of ~600 TFs were downloaded from the cis-BP database (Weirauch et al., 2014). Promoter sequences, comprising 1000 bp upstream of rice genes, were downloaded from AgBase (McCarthy et al., 2006). The promoters were then scanned for at least one or more occurrences of the cis-BP motifs using the FIMO tool in the MEME suite (Bailey et al., 2015). Motifs that were found in more than 50% of all the genes were treated as 'constitutive elements' and removed. Genes harboring all the remaining motifs with a *p*-value < 1E-10 were linked to the corresponding TFs and the motif association network was created for benchmarking.

Benchmarking

For each of the three reference networks, precision (*P*) of the top 500,000 edges was calculated as the ratio of correctly identified edges over total edges predicted, and recall (*R*) was calculated as the ratio of known edges that were correctly predicted over total number of predicted edges. The

overall accuracy score was calculated as the harmonic mean of *P* and *R* (*F* score) as: $F = 2*(P * R)/ P + R$, where $P = TP/ TP + FP$, and $R = TP/TP + FN$.

Evaluation of MCL threshold for clustering

The inflation parameter *I* required for clustering using MCL was evaluated as follows: For clusters obtained at each of *I* values within the range between 1.5 and 4.25 with increments of 0.25, a contingency matrix *T* was computed. Each column in *T* represented a functional class from the GO BP categories and rows represented observed clusters. The cell value T_{ij} represented the number of genes found in common between the i^{th} column and the j^{th} row. Using *T* for every value of *I*, the Sensitivity (*Sn*) and Positive Predictive Value (*PPV*) was computed as

$$Sn = \frac{\sum_{i=1}^{n} Sn_i}{n}, \quad PPV = \frac{\sum_{j=1}^{m} PPV_j}{m}$$

Where $Sn_i = \frac{max\,(T_{ij})}{N_i}$ and $PPV_j = \frac{max\,(T_{ij})}{M_j}$, and N_i corresponds to the size of the BP category and M_j corresponds to the size of the cluster. The geometric accuracy (ACC_g) was then calculated as

$$Acc_g = \sqrt{PPV \cdot Sn}\,.$$

The separation between clusters, indicating how well each complex represents a functional category, was estimated row-wise as

$$Sep_{r_i} = \sum_{j=1}^{m} \left(\frac{T_{ij}}{\sum_{j=1}^{m} T_{ij}} \cdot \frac{T_{ij}}{\sum_{i=1}^{n} T_{ij}}\right)$$

and column-wise as

$$Sep_{c_j} = \sum_{i=1}^{n}\left(\frac{T_{ij}}{\sum_{j=1}^{m} T_{ij}} \cdot \frac{T_{ij}}{\sum_{i=1}^{n} T_{ij}}\right).$$

Then the overall separation was computed as the geometrical mean of row and column separation as

$$Sep = \sqrt{Sep_c \cdot Sep_r}$$

The corresponding values of Acc_g and Sep were then plotted for each value of I in R.

Processing RNA-seq reads and test of differential expression.

Tophat version 2.0.12 was used to align raw paired-end reads from both the drought samples to the *Oryza sativa* cv. Nipponbare release 7 of the MSU Rice Genome Annotation Project. This is the latest release of the reference genome of rice and the General Feature Format (GFF) file contains a total of 55,987 loci as gene models, of which 4665 models were detected to have an alternatively spliced isoform. Correspondence between the replicates was computed by dividing the genome into 10 Kb bins and counting the reads that fell within each bin in each sample. The pairwise correlation values were then evaluated between read coverages and plotted as a scatter plot in R. The number of reads per gene-model were counted using featureCounts (Liao et al., 2014) using 'exons' as features. Genes with 0 read counts across all the samples were removed from the further analyses. EdgeR, Deseq2 and limma as count based methods, and Cufflinks as an FPKM based method were evaluated for their performance in detecting DE. Replicates were not averaged, instead a condition factor was set in the contrast matrix for normalization in the three count based models using all the samples. Genes were then ranked based on their resulting *p/q*

values indicating how confident a method is in declaring a gene as differentially expressed. Performance was then evaluated by comparing the obtained ranks of the *bona fide* drought response genes identified from six independent microarrays.

A custom R script was written to automate the process of counting reads, detecting DE, ranking and performing the aggregation.

Identification differentially expressed genes in microarray data

For identification of DG list, DE was estimated in six drought experiments in rice downloaded from the GEO. Raw data was background corrected, normalized and scaled. Genes with very low variation were filtered based on the IQR range of the sample for reliable detection of DE. DE was tested using the limma model in R (Ritchie et al., 2015). The resulting p values were converted to q values to control for false discovery rate using the qvalue package in R. Genes with q value < 0.01 were declared as significantly differentially expressed (DE) in each dataset. An intersection of DE genes in all the datasets represented the DG list used for evaluating RNA-seq analysis methods.

Derivation of functional categories from online catalogs

The GO graph .obo file was downloaded from the consortium database (http://www.geneontology.org/page/download-ontology). GO annotations of rice were downloaded from the plantGSEA database (Yi et al., 2013). Annotations were propagated from the leaves of the GO hierarchy towards the root, satisfying the parent-child relationships (true path rule) using a custom Perl script. Annotation categories were filtered to retain only those that

possess more than 10 or less than 500 genes for enrichment analysis. Redundant terms were identified as those that had an overlap of more than 90% in their corresponding annotations, and the term with lesser number of annotations was removed. The KEGG pathways mappings for rice was downloaded from the plantGSEA database, and filtered for pathways that had a large number of genes (first three terms in the hierarchy).

Geneset enrichment analysis

For linking clusters to GO BP terms, KEGG pathways and CREs, overlaps between a given cluster and a given category were calculated and the significance of the observed overlap was determined used the cumulative hypergeometric test. The resulting *p* values were corrected for multiple testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995), and -1 log (*q* value) was reported as enrichment scores. For enrichment analysis of clusters in transcriptomes, PAGE model (Kim and Volsky, 2005) was implemented and $\log_2$ of fold change values were used. The *p* value of enrichment was calculated under the standard normal distribution for clusters larger than size 10 and smaller than size 1000.

References

Bailey TL, Johnson J, Grant CE, Noble WS (2015) The MEME Suite. Nucleic Acids Research **43:** W39-W49

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucl Acids Res **35**

Basu S, Krishnan A, Ambavaram M, Rahman L, Ramegowda V, Pereira A (2014) Identification of genes directly regulated by a transcription factor in rice. Prot. Exchange doi:10.1038/protex.2014.039. http://www.nature.com/protocolexchange/protocols/3423

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57:** 289-300

Castillo-Davis CI, Hartl DL (2003) GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics **19:** 891-892

Childs KL, Davidson RM, Buell CR (2011) Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. PLoS ONE **6:** e22196

D'Haeseleer P (2005) How does gene expression clustering work? Nat Biotech **23:** 1499-1501

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

Ficklin SP, Luo F, Feltus FA (2010) The Association of Multiple Interacting Genes with Specific Phenotypes in Rice Using Gene Coexpression Networks. Plant Physiology **154:** 13-24

Gordân R, Hartemink AJ, Bulyk ML (2009) Distinguishing direct versus indirect transcription factor–DNA interactions. Genome Research **19:** 2090-2100

Gu H, Zhu P, Jiao Y, Meng Y, Chen M (2011) PRIN: a predicted rice interactome network. BMC Bioinformatics **12:** 161

Huttenhower C, Hibbs M, Myers C, Troyanskaya OG (2006) A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics **22:** 2890-2897

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE **5:** e12776

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249-264

Iwasaki T, Yamaguchi-Shinozaki K, Shinozaki K (1995) Identification of a cis-regulatory region of a gene in Arabidopsis thaliana whose induction by dehydration is mediated by abscisic acid and requires protein synthesis. Mol Gen Genet **247:** 391-398

Jan A, Maruyama K, Todaka D, Kidokoro S, Abo M, Yoshimura E, Shinozaki K, Nakashima K, Yamaguchi-Shinozaki K (2013) OsTZF1, a CCCH-Tandem Zinc Finger Protein, Confers Delayed Senescence and Stress Tolerance in Rice by Regulating Stress-Related Genes. Plant Physiology **161:** 1202-1216

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

Joshi A, De Smet R, Marchal K, Van de Peer Y, Michoel T (2009) Module networks revisited: computational assessment and prioritization of model predictions. Bioinformatics **25:** 490-496

Joshi R, Wani SH, Singh B, Bohra A, Dar ZA, Lone AA, Pareek A, Singla-Pareek SL (2016) Transcription Factors and Plants Response to Drought Stress: Current Understanding and Future Directions. Frontiers in Plant Science **7:** 1029

Jung KH, Cao P, Seo YS, Dardick C, Ronald PC (2010) The Rice Kinase Phylogenomics Database: a guide for systematic analysis of the rice kinase super-family. Trends Plant Sci **15:** 595-599

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

Lata C, Yadav A, Prasad M (2011) Role of plant transcription factors in abiotic stress tolerance. *In*: Abiotic Stress Response in Plants-Physiological, Biochemical and Genetic Perspectives. InTech.

Lee T, Oh T, Yang S, Shin J, Hwang S, Kim CY, Kim H, Shim H, Shim JE, Ronald PC, Lee I (2015) RiceNet v2: an improved network prioritization server for rice genes. Nucleic Acids Res **43:** W122-127

Lee T-H, Kim Y-K, Pham TTM, Song SI, Kim J-K, Kang KY, An G, Jung K-H, Galbraith DW, Kim M, Yoon U-H, 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 Physiology **151:** 16-33

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30:** 923-930

Li, Z. K., Pinson, S. R., Stansel, J. W., and Paterson, A. H. (1998) Genetic dissection of the source-sink relationship affecting fecundity and yield in rice (Oryza sativa L.). Mol. Breed. 4, 419–426

Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics **26:** 2347-2348

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology **15:** 550

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. Nat Meth **9:** 796-804

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics **7:** S7-S7

McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res **40:** 4288-4297

McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB, Barrell DG, Hill DP, Dolan ME, Williams WP, Luthe DS, Bridges SM, Burgess SC (2006) AgBase: a functional genomics resource for agriculture. BMC Genomics **7:** 229

Meyer PE, Lafitte F, Bontempi G (2008) minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. BMC Bioinformatics **9:** 461

Priya P, Jain M (2013) RiceSRTFDB: A database of rice transcription factors containing comprehensive expression, cis-regulatory element and mutant information to facilitate gene function analysis. Database: The Journal of Biological Databases and Curation **2013:** bat027

Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T (2014) deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res **42:** W187-191

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research **43:** e47-e47

Sato Y, Namiki N, Takehisa H, Kamatsuki K, Minami H, Ikawa H, Ohyanagi H, Sugimoto K, Itoh J-I, Antonio BA, Nagamura Y (2012) RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Research

Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics **18:** S231-S240

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research **43:** D447-D452

Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R, Handakumbura PP, Xiong G, Wang C, Corwin J, Tsoukalas A, Zhang L, Ware D, Pauly M, Kliebenstein DJ, Dehesh K, Tagkopoulos I, Breton G, Pruneda-Paz JL, Ahnert SE, Kay SA, Hazen SP, Brady SM

(2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature **517:** 571-575

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotech **31:** 46-53

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

van Dongen S, Abreu-Goodger C (2012) Using MCL to extract clusters from networks. Methods Mol Biol **804:** 281-295

Vermeirssen V, De Clercq I, Van Parys T, Van Breusegem F, Van de Peer Y (2014) Arabidopsis Ensemble Reverse-Engineered Gene Regulatory Network Discloses Interconnected Transcription Factors in Oxidative Stress. The Plant Cell **26:** 4656-4679

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJ, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell **158:** 1431-1443

Wilson ZA, Zhang D-B (2009) From Arabidopsis to rice: pathways in pollen development. Journal of Experimental Botany **60:** 1479-1492

Yang T, Poovaiah BW (2002) A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. J Biol Chem **277:** 45049-45058

Yang X, Yang Y-N, Xue L-J, Zou M-J, Liu J-Y, Chen F, Xue H-W (2011) Rice ABI5-Like1 Regulates Abscisic Acid and Auxin Responses by Affecting the Expression of ABRE-Containing Genes. Plant Physiology **156:** 1397-1409

Yi X, Du Z, Su Z (2013) PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res **41:** W98-103

Yilmaz A, Nishiyama MY, Fuentes BG, Souza GM, Janies D, Gray J, Grotewold E (2009) GRASSIUS: A Platform for Comparative Regulatory Genomics across the Grasses. Plant Physiology **149:** 171-180

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

Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics **20:** 3594-3603

Chapter 3: SANe: The Seed Active Network For Mining Transcriptional Regulatory Programs of Seed Development in Arabidopsis.

Abstract

Seed development is an evolutionarily important phase of the plant life cycle that governs the fate of next progeny. Distinct sub-regions within seeds have diverse roles in protecting and nourishing the embryo as it enlarges, and for the synthesis of storage reserves that serve as an important source of nutrients and energy for germination. Several studies have revealed that transcription factors (TFs) act in fine coordination to regulate target genes that ensure proper maintenance, metabolism, and development of the embryo. Here, we present genome-wide predictions of seed-specific regulatory interactions between TFs and their target genes in the model plant *Arabidopsis thaliana*. The network is based on a panel of high-resolution seed-specific gene expression dataset and takes the form of a module-regulatory network. TFs that are well studied in the literature were often found at the top of the predicted ranks for the module that corresponds to their validated function role. Furthermore, we brought together a dedicated web resource for a systematic analysis of transcriptional-level regulatory programs underlying the development of seeds (https://plantstress-pereira.uark.edu/SANe/). The platform will enable biologists to query a subset of modules, TFs of interest, as well as analyze new transcriptomes to find modules significantly perturbed in their experiment.

Introduction

The evolutionary success of plants lies in their ability to produce seeds and their dispersal, which facilitates the progression of generation. Seeds are complex structures that help plants halt their life cycle under unfavorable conditions and resume growth once the environmental conditions become favorable. Like all angiosperms, in Arabidopsis, a double fertilization event marks the beginning of seed development that progresses into the development of embryo, endosperm and seed coat over a period of 20-21 days after pollination. These morphologically distinct sub-compartments within a seed play diverse roles and function in concert during the entire phase of seed formation. During maturation, synthesis of storage reserves occur and traits like desiccation tolerance and dormancy are acquired. These seed storage reserves fuel for seedling emergence during germination.

Several transcription factors (TFs) that regulate various aspects of seed development as well as germination have been revealed by genetic screens (Grossniklaus et al., 1998; Lotan et al., 1998; Ogas et al., 1999; Johnson et al., 2002; To et al., 2006). Among these TFs, three members of the B3 super family, namely, *LEAFY COTYLEDON 2 (LEC2)*, *ABSCISIC ACID INDENSITIVE 3 (ABI3)* and *FUSCA3 (FUS3)*, along with two members of the LEC1-type, *LEC1* and *LEC1-LIKE*, that together form the 'LAFL' network (Jia et al., 2013), are the most prominent players of seed maturation. However, the existing LAFL network is still incomplete and represents only a subset of regulatory networks active during seed development. The functional roles of several other TFs that express in seed tissues remains largely unknown. Although genetic interactions, functional redundancy and cooperativity between TFs will be more accurately revealed by genetic perturbations, an underpinning of seed regulatory networks from a computational standpoint will provide tools for quick identification and prioritization of candidates for experimentation *in vivo*.

DNA microarrays have served as efficient experimental systems for simultaneously probing genome-wide transcriptional level activities of specific cellular states. In recent years, an upsurge in the availability of these high-throughput gene expression datasets motivated coexpression based approaches applied to understanding gene function. An integrative analysis of expression datasets enables estimating similarity in patterns of gene expression across a diverse set of experimental conditions. Genes with similar expression profiles are grouped into clusters of coexpressed genes. Functional (Castillo-Davis and Hartl, 2003) and genomic (Huttenhower et al., 2009) annotations of these gene clusters then aid in making functional predictions of uncharacterized genes within these clusters (Childs et al., 2011). There are several such coexpression databases across many model organisms that are now being actively used in gene function prediction and gene prioritization for experimental essays in plants (Obayashi and Kinoshita, 2011; Sato et al., 2012; Yim et al., 2013; Aoki et al., 2016).

Coexpression networks, however, lack information about regulatory interactions encoded in the expression data. Genes encoding regulatory proteins (e.g., TFs) coordinately regulate the biological functions of multiple target genes by directly interacting with their promoters and activating or repressing their expression. Since TFs are themselves transcriptionally regulated, they can also be targets of other TFs, giving the network a hierarchical structure (Ma et al., 2004; Spitz and Furlong, 2012). Hence, a strongly coexpressed TF-gene pair might not necessarily mean a direct physical interaction, but can be observed as an indirect regulatory effect, even if they co-occur in a single functionally related cluster. Moreover, the affinity of a TF for a target gene can be highly tissue-specific or according to the metabolic needs of the cell. Therefore, to deduce a regulatory network prioritizing TFs, the underlying expression data should have a unifying biological context (e.g. datasets for a specific tissue or condition) and coexpressed edges should

be filtered for indirect interactions to minimize false positives. However, inferring accurate regulatory networks using solely gene expression data requires a large number of empirical data points for each space and time combination, for a robust statistical and biological inference. Nonetheless, for plant biologists, accumulated datasets in Arabidopsis are large enough to elucidate specificity of coexpression and predict key functional roles of TFs.

In recent years, several reverse engineering solutions have been brought forward that aim to model coexpression data in a way such that direct interactions involving known regulatory genes are given a priority (Basso et al., 2005; Faith et al., 2007; Huynh-Thu et al., 2010). These algorithms use a successive edge filtering step to recover potentially direct interactions between TFs and their targets. For example, the ARACNE algorithm assumes that in a triplet of connected nodes, the edge with lowest coexpression score is representative of an indirect interaction (Margolin et al., 2006). The GENIE method sets a feature selection problem for every gene to find the best subset of regulators from all the remaining genes (Huynh-Thu et al., 2010). The CLR algorithm aims to identify direct transcriptional interactions by using a background correction scheme that suppresses noise arising due to high correlations between indirect interactions (Faith et al., 2007). These algorithms have been successfully used for inferring plant gene regulatory networks (Yu et al., 2011; Chavez Montes et al., 2014; Vermeirssen et al., 2014).

In the work presented here, we focused on a recently published gene expression dataset arising from the seed development phases of Arabidopsis (Belmonte et al., 2013), and dissected a regulatory network highly predictive of seed-specific functions of TFs (Fig. 3.1). First, we harnessed the power of coexpression and graph clustering to partition genes into functionally related modules and mapped the spatio-temporal activities of these modules. Simultaneously, for every identified TF in the Arabidopsis genome, we computed its partial coexpression score with

every possible target gene and used these scores as a parameter for gene set enrichment analysis using coexpressed modules as gene sets. In this way, we could identify the modules that were statistically-most-likely targets of each TF. Using systematic reduction of data points and prior knowledge from the literature to interpret the associations, we observed that several TFs that are known to have an aberrant seed phenotype were predicted as the top regulators of modules for which their function has been experimentally validated. For example, a recently discovered association between the TF *AGL67* and desiccation tolerance (González-Morales et al., 2016), and *MYB107* and suberin (Lashbrooke et al., 2016) was correctly predicted in our network. These and several other correctly predicted associations (described later in the text) motivated us to create an online resource for the community. Our network, which we termed the 'Seed Active Network' or SANe, is hosted at https://plantstress-pereira.uark.edu/SANe/ to provide a network-based understanding of seed development.

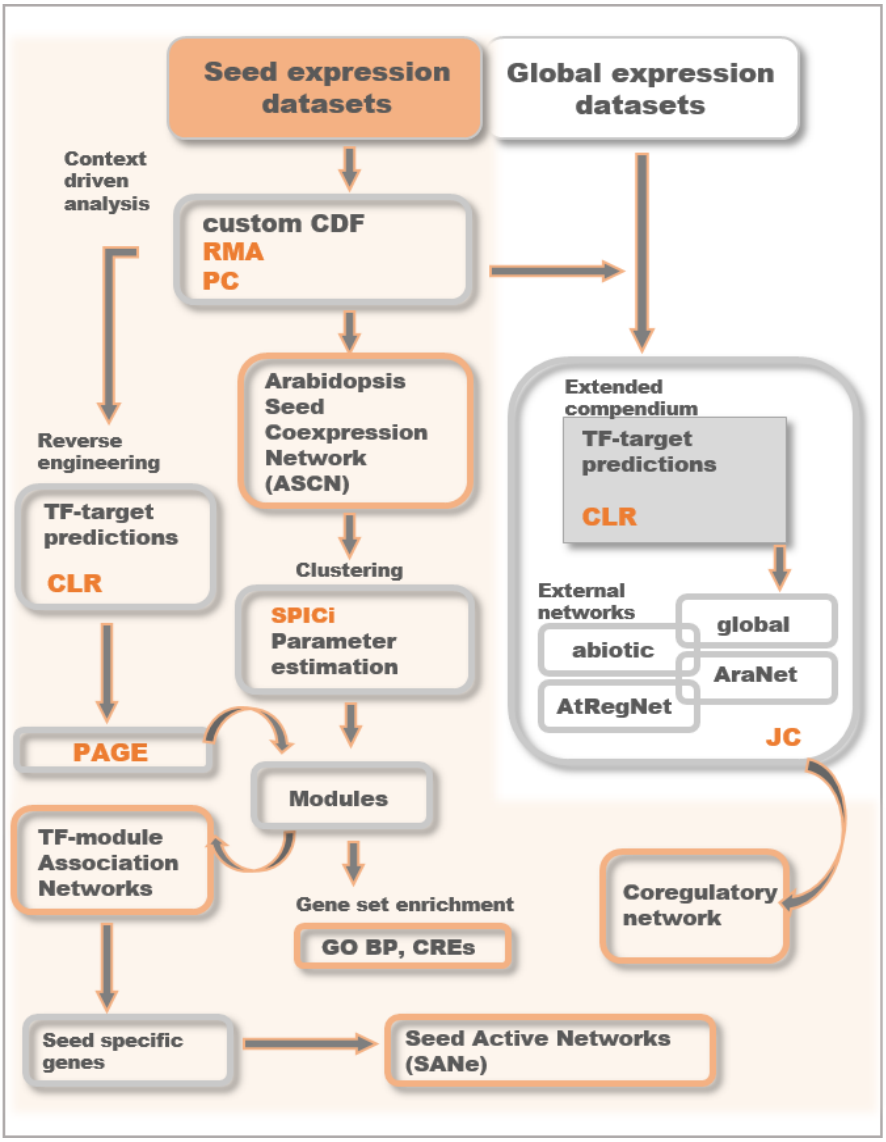

Figure 3.1: Pipeline for tissue-specific module regulatory network analysis. Two separate Arabidopsis gene expression compendiums (EC) were created: one from a seed-specific expression data series (GSE12404) and one from non-tissue-specific (global) 140 expression datasets. Datasets in both the EC were normalized individually using RMA algorithm in R. Z scores of Pearson's Correlations (PC) were calculated for all gene-pairs in both the EC. From seed EC, gene pairs with PC >0.73 (Z > 1.96) were connected to create the Arabidopsis Seed coexpression network (ASCN). ASCN was then clustered using SPICi at a range of clustering thresholds *(T_d)*, and an optimal clustering parameter was chosen based on genome coverage and coherence of genes as a functional group. 1563 clusters obtained at T_d 0.80 were tested for enrichment of biological processes from the gene ontology and known plant *cis* regulatory elements for multiple databases. A list of 1921 TFs was supplied to the CLR (context likelihood of relatedness) algorithm to predict their targets in both the EC. In the seed EC, the PAGE algorithm was used to score the enrichment of CLR-weighted targets in the ASCN clusters, and a TF-module association network was created. The network was queried with a list of genes

expressed predominantly in the seed as compared to other organs/tissues, and the Seed Active Network or SANe, was derived. Simultaneously, the seed-specific network was compared with the network created using the global EC and multiple other Arabidopsis regulatory networks downloaded from published studies.

Results

Seed coexpression network

To avoid implementing procedures of minimizing batch effects and the errors associated with microarray data integration (Chen et al., 2011; Nygaard et al., 2015), we chose Arabidopsis gene expression profiles from the data super series labeled GSE12404 in the gene expression omnibus (GEO) database (Barrett et al., 2007). This series is comprised of 87 samples derived from 6 discrete stages of seed development, and 5-6 different compartments within each stage, reflecting the most comprehensive source of Arabidopsis seed-specific gene expression profiles. With a sample size large enough for statistical inferences, these datasets were also devoid of the ambiguities regarding due to the context under which the experiment was performed (intra-laboratory bias), one of the major problems in context-driven integrative analyses of gene expression data. We normalized and summarized this expression data into an integrated gene expression matrix using a custom CDF file of Arabidopsis microarray to reduce off-target hybridizations (Harb et al., 2010). Pearson's correlations (PC) scores between all gene-pairs in the gene expression matrix were then calculated and mapped to Z scores using Fisher's Z-transformation (Huttenhower et al., 2006). Gene pairs with significantly high correlation in expression (PC 0.753, Z-score >1.96) were connected and the rest filtered. We named this core of raw coexpression data with ~7.6 million edges as the Arabidopsis seed coexpression network (ASCN).

Identification of clusters in coexpression data

Identification of communities, or clustering, is the most prominent step in network based interpretation of genomic data. In terms of gene expression data, clustering provides a useful way

to group genes with similar expression profiles together. The need for gene grouping is based on the percept that expression similarity is indicative of similarity in function (Eisen et al., 1998). Therefore, clustering furthers an understanding of the function of a previously uncharacterized gene, based on known functions of other members of the same group. However, the choice of clustering method heavily influences the accuracy of functional predictions (Yeung et al., 2001). Clustering algorithms typically require either a predefined number of clusters, as in k-means clustering, or the process is semiautomatic (Langfelder and Horvath, 2008), and is sometimes computationally expensive.

In our network framework, we used an unbiased data-driven method to cluster genes within the ASCN. The density of a cluster, measured as the ratio of the number of observed edges in a cluster to the total number of expected edges, reflects cohesiveness among the members of the same cluster. The SPICi algorithm evaluates density to group similar genes in a biological network, while considering the confidence weight on each edge (Jiang and Singh, 2010). We sought to identify an optimum density threshold (T_d) that yields clusters at a granularity that delivers biological information, while preserving the inherent topological features of the network. A range of T_d values were evaluated for performance in loss or gain of information, with a goal of separating genes into as many clusters as possible, without losing many genes originally present on the microarray. At T_d 0.80, 84% of the ASCN genes formed 1563 clusters, after which a significant loss of information occurred, as indicated by a sharp fall in the fraction of total genes retained (Fig. 3.2 A). At the same threshold of 0.80, the average modularity within clusters was also maximized (at a bearable cost of gene loss) (Fig 3.2 B). Modularity measures how functionally separable the clusters are, in the sense that how well genes within a clusters interact with each other as compared to genes outside the cluster (Albert, 2005).

For a function-level analysis, it is also important that genes within each cluster are representative of common biological functions, as grouping genes would not yield any functional predictions if at least one putative function of the group is not known. To further establish confidence in T_d 0.80 as the best solution for partitioning, we evaluated each T_d for its ability to categorize known information about Arabidopsis biological pathways derived from the Gene Ontology (GO) annotated gene sets in the biological process (BP) category. Full set of annotation terms satisfying the parent-child relationships were used to find overlaps with clusters obtained at every T_d. The significance of overlap was tested under the hypergeometric distribution (see "Methods"). The functional coherence of the network, evaluated based on the total number of clusters with enriched BP terms, total number of distinct BP terms and the overall functional enrichment score, was also found to be best preserved at T_d 0.80 (Fig 3.2 B and 3.2 C).

Overall, the network lost its stability and collapsed at T_d values exceeding 0.80, as indicated by all measured clustering parameters (Fig. 3.2). Hence, 1563 dense clusters obtained at T_d 0.80 were used for further analysis. The total number of genes in these modules amounts to 17,949 (Supplemental Table S3.1).

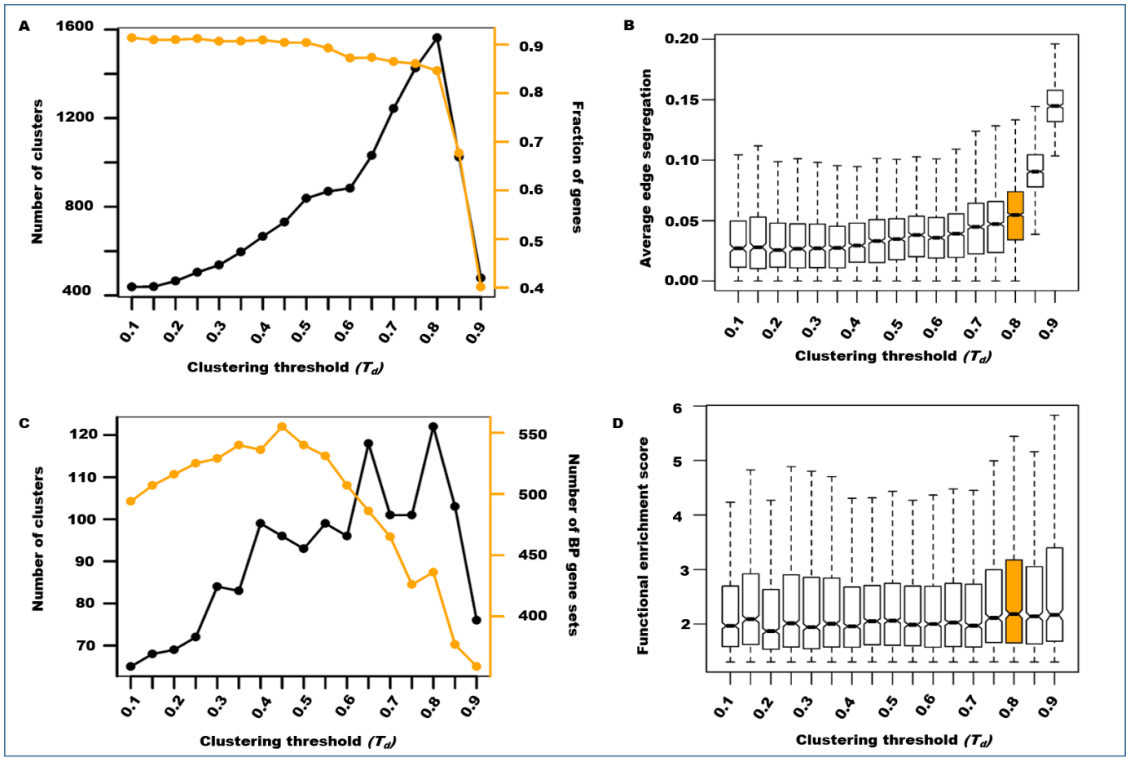

Figure 3.2: Evaluation of clustering threshold (T_d). Genes from the Arabidopsis seed coexpression network were clustered at a range of T_d values shown on the X axis of all the figures. Each T_d was examined by: A) A genome coverage plot measuring the number of clusters yielded and the fraction of original genes retained (orange line corresponding orange Y axis). B) Boxplots showing average edge segregation of all the clusters, indicating overall modularity of the network within each T_d. C) A plot showing number of clusters enriched with atleast one BP term and the total number of BP terms retained (orange line corresponding orange Y axis) and D) Boxplots summarizing the enrichment scores [-1*log(FDR)] of the hypergeometric p-values obtained by BP-cluster overlap analysis.

Transcriptional regulators of seed modules

Modules of coexpressed genes in ASCN retained information about possible functional interactions between genes and their responses during different stages of seed development. This greatly expanded upon the currently available functional annotations of Arabidopsis genes, as the genes that were lacking functional annotations now have atleast one putative function assigned based on their module participation. The next task was to leverage on this information in the coexpression data and identify key TFs that statistically associate with each of the ASCN modules. There are 1921 unique locus IDs in the Plant Transcription Factor Database (Jin et al., 2014), the AGRIS database (Yilmaz et al., 2011) and the Database of Arabidopsis Transcription Factors (Guo et al., 2005), corresponding to TF genes in Arabidopsis. We used this comprehensive list to obtain transcriptional regulators for our analysis.

Simply associating genes as targets of TFs that they 'highly coexpress' with (first neighbors) is prone to the occurrence of false positives in a genome-scale analysis. This occurrence is mainly due to correlations arising from indirect regulation or coincidental coexpression of genes involved in different and unrelated processes that need to be active under the same circumstances. To minimize this effect, we calculated how likely a predicted TF-gene interaction was given the empirical background distribution of correlation scores of both the genes under consideration (Faith et al., 2007) (reported as a Z-score, see Methods) (Supplemental Table S3.2). Next, we sought to identify those modules that had higher enrichment of most probable targets for each TF. Instead of choosing an arbitrary cutoff for selecting targets, we used the entire set of predictions for each TF, weighted by Z-scores, and worked under the framework of Parametric Analysis of Gene set Enrichment (PAGE) (Kim and Volsky, 2005). The PAGE algorithm uses the normal distribution for statistical inference and states the degree of enrichment (here 'association') of a

given gene set (here module) amongst the most highly scored predicted targets of a given TF. This analysis is essentially similar to that of a two-tail enrichment test with GO BP terms (treated as gene sets) (Ambavaram et al., 2011). Here, the difference was that gene sets from coexpression clusters observed in a specific tissue was used. To provide a normal distribution for association scoring, we used only those modules that had more than 10 genes, as suggested by the authors of the PAGE algorithm. Using this robust formulation, 1819 TFs were linked to 278 modules comprised of 10,526 genes (cluster 1 with 1621 genes was considered an outlier cluster because it contained disproportional number of genes as compared to other clusters). We labeled this network core as 'TF-Module Network' (TMN). TMN is represented as a matrix with TFs in rows and modules in columns, with each cell in the matrix representing a TF-module association score given by PAGE (Fig. 3.3 A).

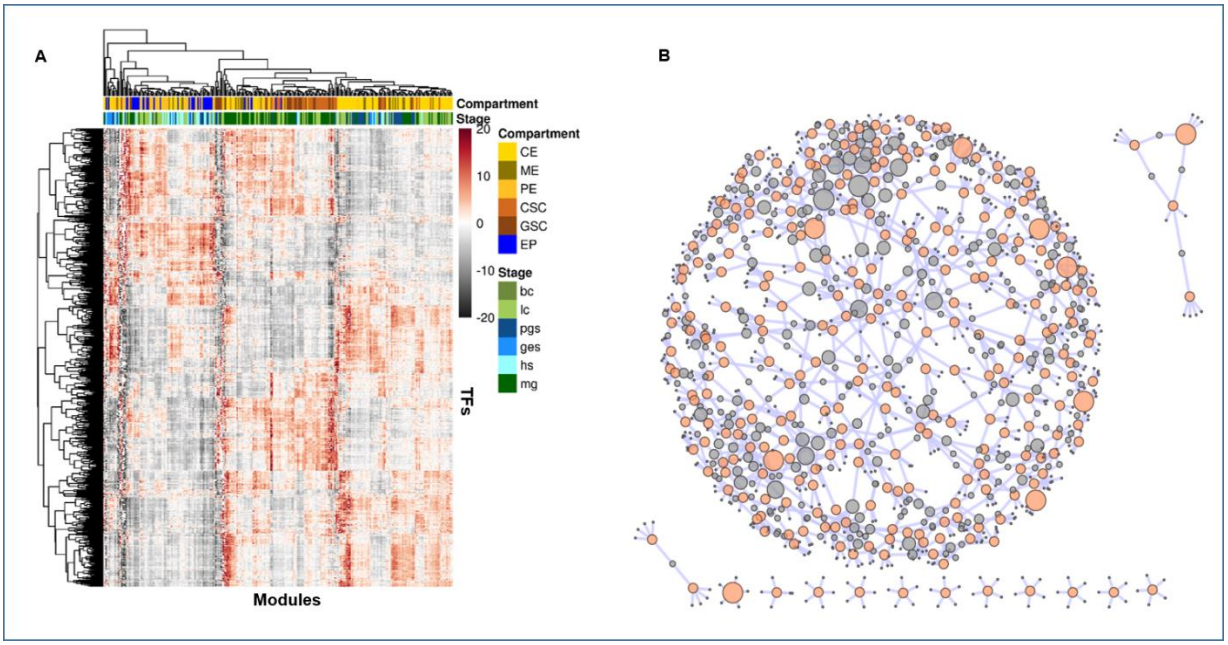

Figure 3.3: TF-Module Association Network (TMN). A) Heatmap representing association scores of 1819 TFs as regulators in the rows, and 10526 genes grouped into 278 coexpression modules represented along the columns. Each grid in the heatmap is color coded according to the level of enrichment of predicted targets of each TF regulator in the corresponding module. The red gradient indicates a positive score and grey indicates a negative score, estimated using the PAGE algorithm. The seed compartment in which the module has maximum expression is color-coded and represented on top of the heatmap (first row), where CE is chalazal endosperm, ME is micropylar endosperm, PE is peripheral endosperm, CSC and GSC is chalazal and general seed coat, respectively, and EP is embryo proper. The development stage in which the module has maximum expression is color-coded and represented on top of the heatmap (second row), where bc is bending cotyledon, lc is linear cotyledon, pgs is pre-globular stage, ges is globular embryo stage, hs is heart stage and mg is mature green stages. B) Predictions for each of the 278 modules were ranked and the top 5 predicted regulators for each module were visualized as a network graph. Each grey circle in the network plot is a TF and each orange circle is a module. The size of the grey circle is proportional to the out-going degree of the TF. Size of the orange circle was set to a constant, except for 9 bigger circles showing the modules described later in the main text. The network was visualized using Cytoscape version 3.3.0. Node names are hidden for ease in visualization. The cytoscape sessions file is provided as supplemental table S3.3, which can be loaded into Cytoscape for node names and further exploration of the network. The heatmap was drawn using gplots package in the R statistical computing environment.

The TMN provides a regulatory map of seed transcriptional activities, in the form of a bipartite graph, with TFs as one set of nodes and sets of genes reduced to their 'functions' as another set of nodes, and edges weighted by the degree of association between the corresponding TF and the function. For visualization, we selected the top 5 predicted TF regulators for each module, ranked based on absolute association scores, and visualized TMN as a graph in Cytoscape (Fig. 3.3B; Supplemental Table S3.3). A total of 900 regulators were represented in top 5 predictions for each of the 278 modules. Most the modules were found indirectly connected due to combinatorial links between their predicted TF regulators, forming a dense network while 11 modules shared no common predicted TF regulators with other modules.

Modules active during seed development

Seed-specific genes were previously discovered as those that were present only in seed tissues, and not in other reproductive or vegetative parts of the plant (Le et al., 2010; Belmonte et al., 2013). We sought for those modules that harbored at least one such gene and identified a core set of 120 modules comprised of 7414 genes. We called these modules as 'active modules'. We reasoned that because these modules retained genes specific to seed development, their coexpression neighborhood – along with the top ranked regulators – will pave way to identification of transcriptional networks modulated specifically during seed development, or involved in important seed functions. Therefore, novel TFs that are already part of these modules, or emerge as the top regulators will automatically become the primary candidates for testing seed phenotypes, largely reducing the search space. Also, the strategy of probing TMN with a list of genes already prioritized had less chances of observing false positives from a gamut of predicted regulatory programs, while making the process of interpreting the regulation patterns easier. We labelled this

core of 120 active modules along with their scored TF regulators as the 'Seed Active Network' (SANe) (Supplemental Table S3.4).

We simultaneously mapped the expression patterns of each module spatially and temporally (seed compartment wise and development stage wise), by averaging the expression of module genes in each seed-compartment irrespective of the development stage or within each development stage irrespective of the seed compartment. After interfacing the expression patterns of each module with BPs and known *cis* regulatory elements (CREs) (Supplemental Table S3.5 and S6; see "Methods") and predicted sets of top regulators, a few modules that had high expression in different seed compartments (embryo, endosperm and seed-coat regions) were visually examined using heatmaps (Fig. 3.4). These modules expand a wide variety of cellular processes, including flavonoid metabolism during seed coat formation, lipid storage and photosynthesis during endosperm development and auxin transport and tissue development from early to late stages of embryogenesis. Visualization of a few modules revealed that there is a high intra-module connectivity between modules that participate in the same developmental program in a tissue-specific manner, albeit with different biological goals (Fig. 3.5). A few such modules are described below.

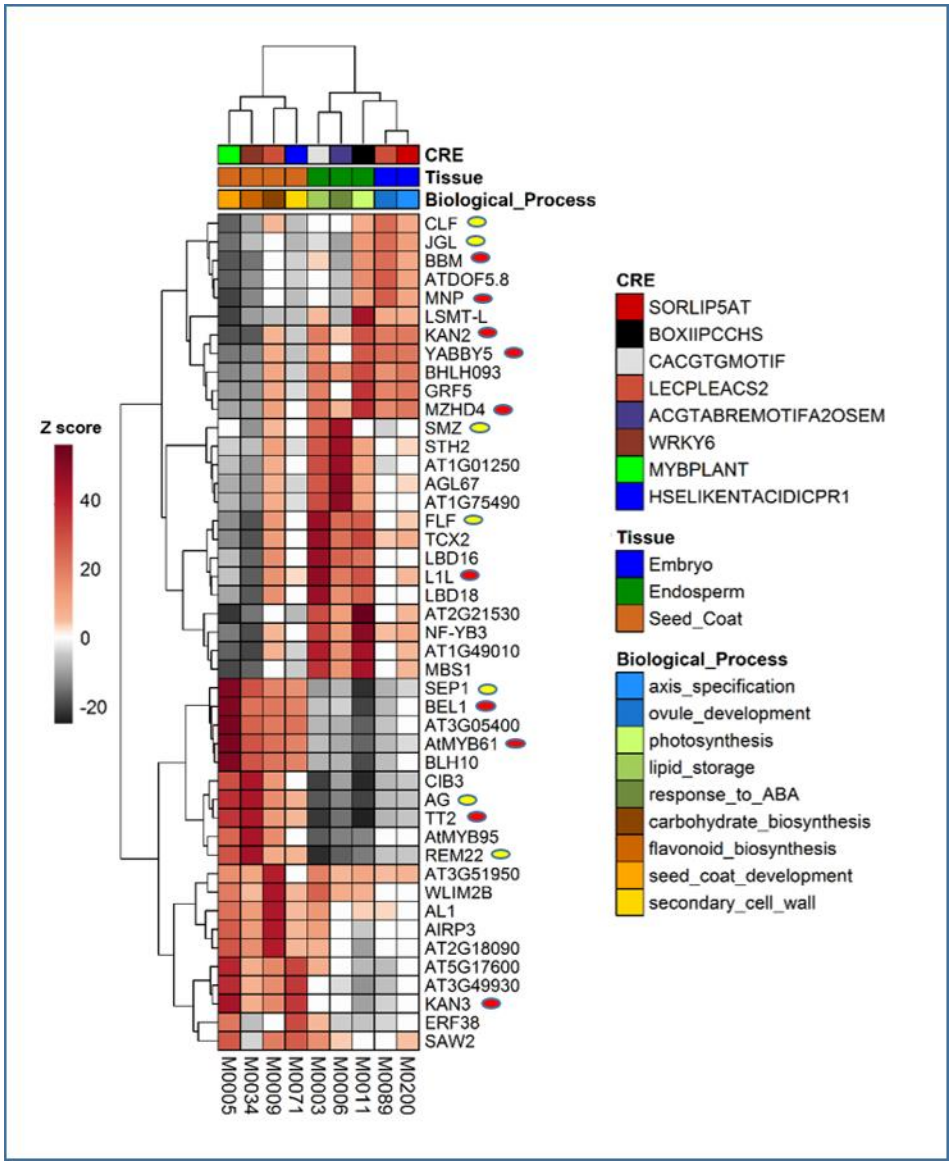

Figure 3.4: Heatmap representation of a subset of the SANe. Modules with high relative expression in embryo, endosperm and seed coat regions were extracted from SANe. Modules are shown in columns and for each module, the top 5 predicted TF regulators are shown in rows. Each grid in the heatmap is colored according to the association score estimated using the PAGE algorithm. Positive and negative scores are shaded in red or black gradient, respectively, as indicated by the color key. Literature identified TFs with validated seed-specific phenotypes or phenotypes observed in other reproductive stages/tissues are marked with a red ellipse or a yellow ellipse, respectively. CRE, cell-type and functional annotations for each module are shown above the heatmap (top three rows; colored boxes). Modules annotated for embryo, endosperm and seed coat are indicated in blue, green and brown boxes, respectively, in the middle row. CRE and functional annotation for each module is color- coded uniquely in the top and bottom rows, respectively.

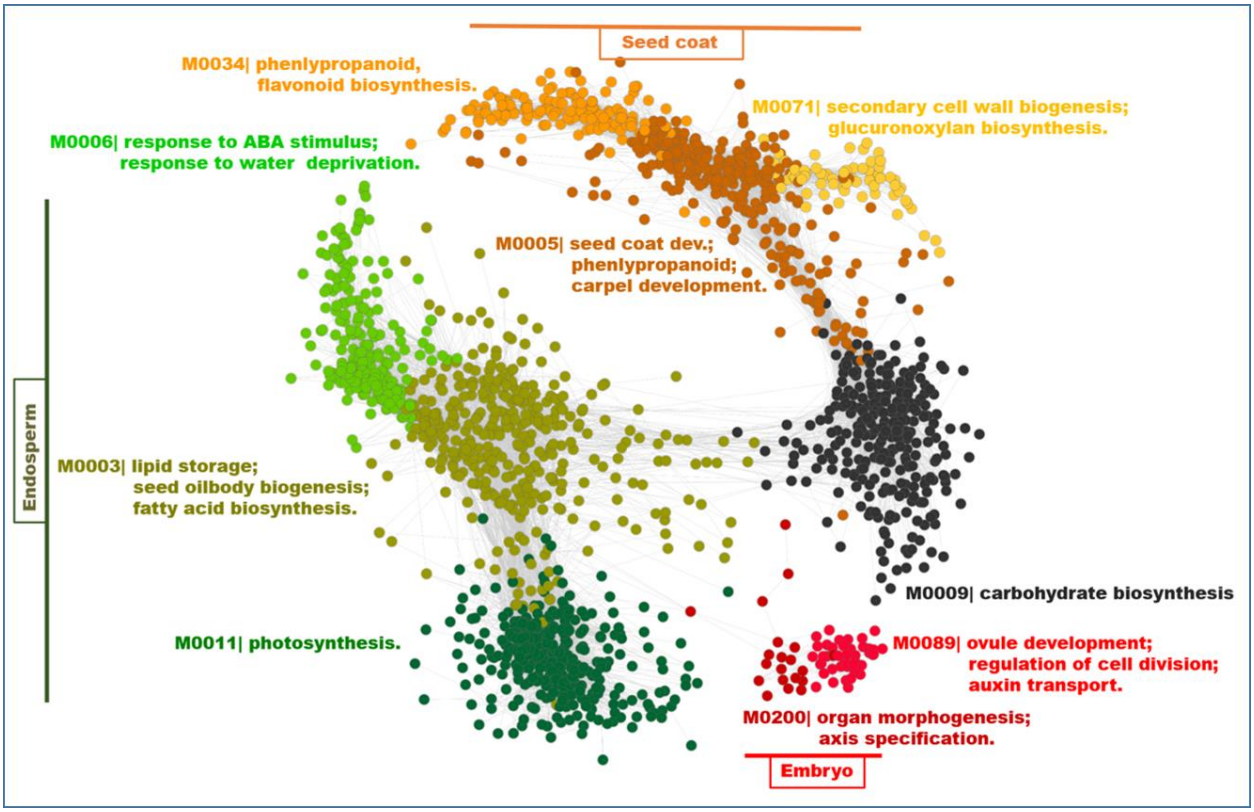

Figure 3.5: Visualization of seed modules. A graphical representation of seed coexpression modules. Each circle represents a gene. Each module is color coded uniquely. Modules are grouped according to the seed compartments (indicated by horizontal or vertical lines and text boxes), and labelled with the BP term most highly over-represented within each module. Genes are left unlabeled to facilitate visualization. The network was drawn in Cytoscape version 3.3.0.

Modules for early embryo development

Three modules designated as M0089, M0200 and M0277 comprised 54, 31 and 33 genes, respectively, expressed at relatively high levels in the embryonic tissue when compared to other seed compartments (Fig. 3.6A). These genes are significantly enriched with BP terms like “organ development”, “tissue development”, “axis specification” and “auxin transport”. This is consistent with processes related to embryo development, involving morphogenesis-related and other cellular processes that govern gene activity related to cell division and expansion, maintenance of meristems and cell fate determination (Wendrich and Weijers, 2013).

M0089 harbors genes related to reproductive tissue development and cell division. ATDOF5.8 (AT5G66940) was predicted as the top regulator of M0089. The *ATDOF5.8* gene is most highly expressed in embryo and meristem cells (Supplemental Fig. S3.1A) based on the Genevisible tool in GENEVESTIGATOR (Zimmermann et al., 2004). It has been shown that ATFOD5.8 is an abiotic stress-related TF that acts upstream of ANAC069/NTM2 (AT4G01550) (He et al., 2015). Interestingly, the *NTM2* gene resides at a locus adjacent to another NAC domain TF, NTM1 (AT4G01540), a regulator of cell division in vegetative tissues (Kim et al., 2006). Kim et al. did not detect *NTM2* expression in leaves by RT-PCR. However, they indicated that because both *NTM* genes have similar structural organization, encoding proteins with a few differences in the protein chain, NTM2 could be involved in similar processes in other tissues. Our predictions suggest that *NTM2* could be in the ATDOF5.8 regulon associated with modulating cell division activity in the seed. This leads to a new testable hypothesis pertaining to regulation of cell division during embryogenesis. Among other known regulators, BABY BOOM (BBM, AT5G17430) was predicted as one of the top ranked TF (rank 4) of M0089. BBM is an AP2 TF that regulates the embryonic phase of development (Boutilier et al., 2002).

YAB5 (AT2G26580) and ATMYB62 (AT1G68320) were predicted the top ranked regulators of M0200 and M0277, respectively. While the agreement of YAB5 as a determinant of abaxial leaf polarity (Husbands et al., 2015) and enrichment of M0200 with GO BP term "axis specification" (GO:0009798) justifies this association, the association of ATMYB62 with M0277 indicates a hormonal interaction likely representing a transition between the growth stages. *ATMYB62* encodes a regulator of gibberellic acid biosynthesis (Devaiah et al., 2009) and is expressed specifically during seed development (Belmonte et al., 2013). M0277 is enriched with "auxin transport" genes (GO:0009926). The *ATMYB62* gene is preferentially expressed in the abscission zone and other reproductive tissues (Supplemental Fig. S3.1B).

Modules for Endosperm Development

The endosperm has a profound influence on seed development by supplying nutrients to the growing embryo (Portereiko et al., 2006; Chen et al., 2015). The importance of endosperm cellularization for embryo vitality has been shown through mutants deficient in endosperm-specific fertilization events (Kohler et al., 2003). The overall seed size depends on endosperm development and is controlled through the relative dosage of accumulated paternal and maternal alleles (Luo et al., 2005).

We found that genes in modules M0003 and M0011 had maximal expression levels in endosperm tissues (Fig. 3.6B). M0003 is significantly enriched with genes involved in lipid storage (GO:0019915) and fatty acid biosynthesis (GO:0006633). LEC1-LIKE (L1L, AT5G47670) emerged as the top regulator of this module. LIL is related to LEAFY COTYLEDON 1 (LEC1) and functions during early seed filling as a positive regulator of seed storage compound accumulation (Kwong et al., 2003). Interestingly, L1L is also part of this module indicating that,

apart from being a master regulator, its activity is also modulated during the late seed filling stages as observed previously (Kwong et al., 2003), which correlates with the overall expression pattern of genes within this module (Supplemental Fig. S3.2). The presence of 44 other TFs in this module, including FUS3 and ABI3, key regulators of seed maturation (Keith et al., 1994; Luerßen et al., 1998; Yamamoto et al., 2009), points to the importance of this module in nutrient supply to the developing embryo. LDB18 (AT2G45420) is a LOB-domain containing protein of unknown function predicted as the second ranked regulator of this module. GENEVESTIGATOR analysis showed that both *L1L* and *LDB18* are most highly expressed in the micropylar endosperm (Supplemental Fig. S3.3).

M0011 is comprised of 357 genes including 7 TFs and is characterized by containing genes with high expression levels in the micropylar endosperm (ME) and the peripheral endosperm (PE). GO enrichment analysis showed the highest scores for photosynthesis (GO:0015979) for genes in this module. Close examination of these genes revealed that virtually all aspects associated with chloroplast formation and function were represented, including chloroplast biogenesis and membrane component synthesis, chlorophyll biosynthesis, plastidic gene expression, photosynthetic light harvesting and electron transport chain, ATP production, redox regulation and oxidative stress responses, Calvin cycle and photosynthetic metabolism, metabolite transport, and retrograde signaling. Interestingly, genes encoding photorespiratory enzymes (glycine decarboxylase, glyoxylate reductase, and hydroxypyruvate reductase) were also present in M0011. Developing oilseeds are known to keep extremely high levels of CO_2 that would suppress photorespiration (Goffman et al., 2004), and the implications of expression of these genes on photosynthetic metabolism are not clear.

The presence of mostly photosynthetic genes in M0011 seems also unusual, but the results are consistent with findings of (Belmonte et al., 2013), showing that specific types of endosperm cells are photosynthetic, as they contain differentiated chloroplasts and express photosynthesis-related genes. Fully differentiated embryos at the seed-filling stages and the chlorophyll-containing inner integument ii2 of the seed coat are parts of oilseeds that are also capable of photosynthesis (Belmonte et al., 2013; Sreenivasulu and Wobus, 2013). Although seeds obtain the majority of nutrients maternally, Arabidopsis embryos remain green during seed filling and maintain a functional photosynthesis apparatus similar to that in leaves (Allorent et al., 2015). As part of photoheterotrophic metabolism, photosynthesis provides at least 50% of reductant in oilseed embryos and CO_2 is re-fixed through the Rubisco bypass that helps to increase carbon-use efficiency in developing oilseeds (Ruuska et al., 2004; Schwender et al., 2004; Goffman et al., 2005; Fait et al., 2006). The roles for photosynthesis in ME and PE remain to be investigated and include (i) providing carbon and energy for storage compound accumulation in the endosperm and the embryo and (ii) increasing the availability of oxygen to the endosperm and differentiating, yet-to-be photosynthetic, embryos in a high-CO_2 environment.

CRE analysis revealed the highest number of motifs enriched in the promoters of genes in M0011, suggesting extensive coordination between different regulators. Light-related motifs BOXIIPCCHS (ACGTGGC), IRO2OS (CACGTGG3), IBOXCORENT (GATAAGR) and the ABA-responsive element ACGTABREMOTIFA2OSEM are the most over-represented motifs in this module. The highest ranked regulator of M0011 is a SMAD/FHA domain-containing protein (AT2G21530) that is most highly expressed in the cotyledons (Supplemental Fig. S3.4A). The known seed-specific regulator of oil synthesis and accumulation WRI1 (AT3G54320) was identified as the sixth ranked regulator of this module and is suggested to be predominantly

expressed in the embryo and endosperm (Supplemental Fig. S3.4B). *WRI1* encodes an AP2/ERF-binding protein and *wri1* seeds have about 80% reduction in oil content relative to the wild type seeds (Ruuska et al., 2002). Genetic and molecular analysis revealed that WRI1 functions downstream of LEC1 (Baud et al., 2007). Along with WRI1 itself, six other TFs are part of this module, including AT2G21530, a zinc finger (C2H2) protein (AT3G02970), NF-YB3 (AT4G14540), PLT3 (AT5G10510), GIF1 (AT5G28640) and PLT7 (AT5G65510).

Modules for Seed Coat development

The seed coat has important functions in protecting the embryo from pathogen attack and mechanical stress. The seed coat encases the dormant seed until germination and maintains the dehydrated state by being impermeable to water. M0034 is comprised of 149 genes with the highest expression in general, and specifically in chalazal seed coat relative to other tissues (Fig. 3.6 C). This module is enriched with genes annotated under the GO BP terms "phenylpropanoid biosynthetic process" (GO:0009699) and "flavonoid biosynthesis process" (GO: 0009813). The AP2/B3-like TF AT3G46770 is highly expressed in seed coat (Supplemental Fig. S3.5A) and predicted as the top regulator in this module. B3 domain TFs are well known for functioning during seed development and transition into dormancy in *Arabidopsis* (Suzuki and McCarty, 2008) and, to some extent, their functions are conserved in cereals (Grimault et al., 2015). The seed-coat-specific expression of AT3G46770 is a compelling incentive for testing AT3G46770 mutants for seed-related phenotypes, which to the best of our knowledge, has never been considered. There were 21 other TFs belonging to this module, of which six are part of the MYB family. TRANSPARENT TESTA 2 (TT2), a MYB family regulator of flavonoid synthesis (Nesi et al., 2001), was ranked fourth in our predictions for this module.

M0071 is composed of 77 genes encoding, surprisingly, only 3 TFs, ERF38 (AT2G35700), BEL1-LIKE HOMEODOMAIN 1 (BLH1, AT2G35940) and a C2H2 super family protein (AT3G49930). This module is enriched with genes involved in "xylan metabolic process" (GO:0045491), "cell wall biogenesis" (GO:0009834), and "carbohydrate biosynthetic process" (GO:0016051). KANADI3/KAN3 (AT4G17695) was predicted as the top regulator of this module. KANADI group of functionally redundant TFs (KAN1, 2, and 3) has been shown to play roles in modulating auxin signaling during embryogenesis and organ polarity (Eshed et al., 2004; McAbee et al., 2006; Izhaki and Bowman, 2007). In the case of another KANADI TF, KAN4, encoded by the *ABERRANT TESTA SHAPE* gene, the lack of the KAN4 protein resulted in congenital integument fusion (McAbee et al., 2006). It is reasonable to hypothesize that KAN3 could be acting in a redundant manner with KAN4 to regulate seed coat formation during late stages of maturation, as the expression pattern of *KAN3* is higher in seed coat than in other organs or cell types (Supplemental Fig. S3.5B).

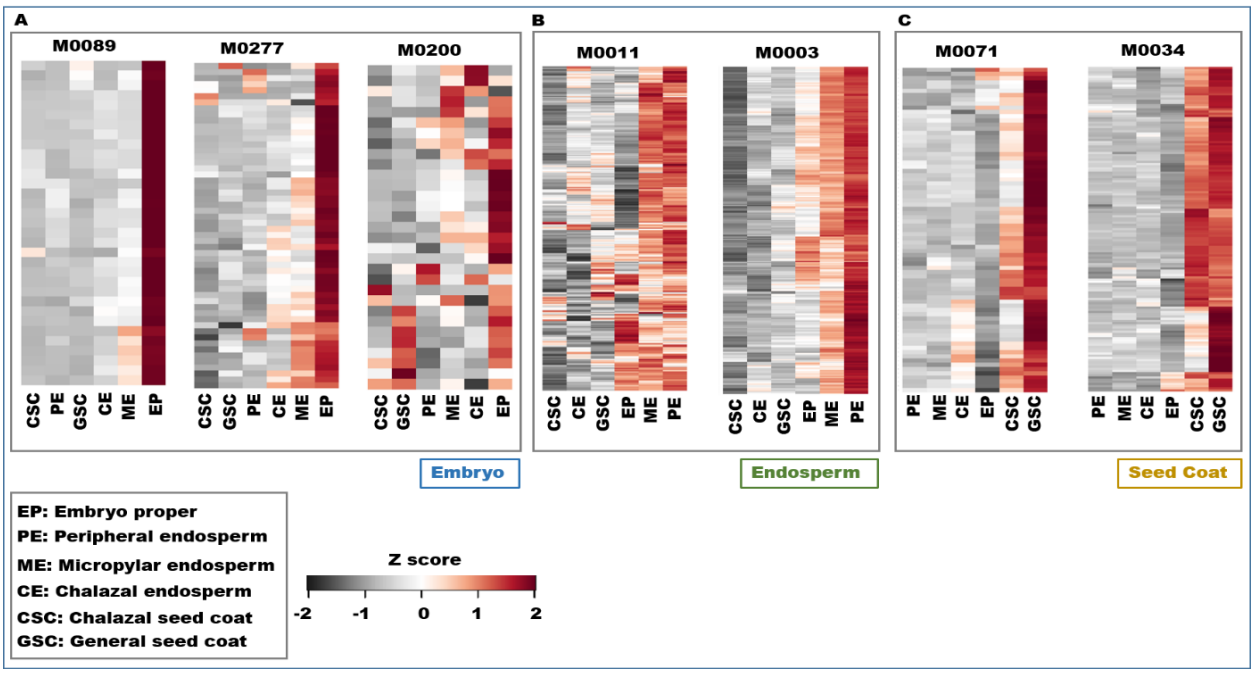

Figure 3.6: Expression profiling of gene modules. Expression patterns of modules in embryo, endosperm and seed coat regions represented as heatmaps in A), B), and C), respectively. Seed compartments are represented as columns and genes as rows. Gene names are hidden for ease in visualization. Expression values of genes in each module were averaged across samples from the same tissue-type/seed-compartment (embryo, endosperm and seed coat). Average expression values were scaled and represented as a Z score in the heatmaps. Red indicates higher expression of a gene in a particular compartment and black gradient indicates lower expression relative to other compartments.

Module M0006 is related to seed desiccation tolerance

M0006 is comprised of 220 genes expressed predominantly during the mature green stage (Fig. 3.7A), and enriched with genes involved in "response to abscisic acid stimulus" (GO:0009737), "response to water" (GO:0009415) and terms related to embryonic development (GO:0009793), altogether suggesting an involvement of these genes in acquisition of desiccation tolerance (DT). We predicted AGL67 (AT1G77950) as a major regulator of this module, among 23 other TFs that are part of this module (Fig. 3.7B). AGL67 has been recently confirmed as a major TF involved in acquisition of DT (González-Morales et al., 2016), validating our prediction. Additionally, the authors of this study analyzed the mutants of 16 genes (TFs and non-TFs) that had reduced germination percentage, of which 12 are in our network and 7 of these are a part of M0006. These 7 genes include PIRL8 (AT4G26050), ERF23 (AT1G01250), OBAP1A (AT1G05510), DREB2D (AT1G75490), AT1G77950 (AGL67), AT2G19320 and MSRB6 (AT4G04840).

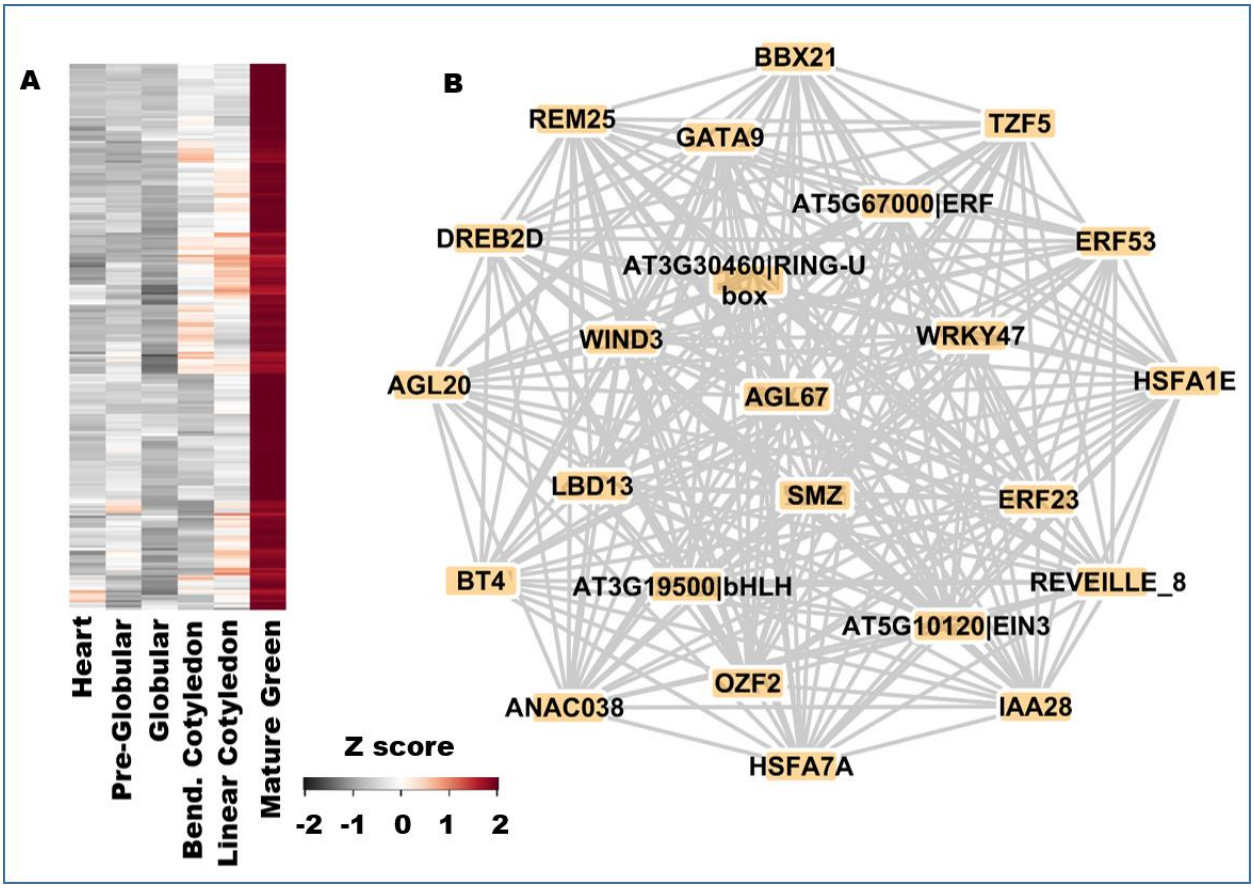

Figure 3.7: Module M0006. A) Expression patterns of genes in module M0006. Seed development stages are represented as columns and genes as rows. Gene names are hidden for ease in visualization. B) Coexpression links between TFs in M0006. Nodes are labeled according to their corresponding gene symbols if present in TAIR, else labeled with their corresponding locus ID and the family the protein belongs to.

Characteristics of seed-specific networks

The primary objective of this network analysis pipeline was to capture gene regulation information in a tissue-specific manner. To examine the effect of this approach and to identify the distinguishing characteristics of the seed regulatory network that differed from a global network (non-tissue specific regulatory network), we extended the seed expression compendium to incorporate an additional set of 140 datasets related to profiling gene expression from various organs of the Arabidopsis plant, including vegetative and seedling growth stages. Using the same reverse engineering approach as described above, we scored each TF-target pair on this extended expression compendium (EEC). Next, to delineate the distinguishing properties of seed networks, we compared the level of coregulation induced by TFs, measured as similarity in the predicted targets of each TF-pair, using Jaccard's coefficient (JC), in both the seed-specific network and the global regulatory network created using EEC. As expected, a larger number of TFs have very few common targets, and this number is high for fewer TFs in both the networks (Fig. 3.8A). A larger number of TFs have similar targets in the seed network at any given JC bin, as compared to the global network.

Although false positives and false negatives are part of any network based predictions, we suspected that the trends observed in comparison of the seed-specific and the global network could be trivial if there were correlated errors arising from the same network prediction pipeline for both networks. To overcome this uncertainty, we downloaded and analyzed the recently published Arabidopsis oxidative stress gene regulatory network predicted from a compendium of microarrays conditioned on abiotic stress (Vermeirssen et al., 2014). This abiotic-stress specific network is essentially a consensus network of an ensemble of reverse engineering algorithms, and performed remarkably well in validations (Vermeirssen et al., 2014). We then computed the

overlaps in the predicted targets of TFs in this network (as done for networks in this study) and observed that it follows a trend very similar to that of the global network (Fig. 3.8A), indicating that there was no major bias introduced by our approach.

To extend the comparisons, we performed the same operation to the *Arabidopsis thaliana* Regulatory Network (AtRegNet) and the AraNet (Lee et al., 2010). AtRegNet harbors about 17,000 direct edges validated for TFs and their target genes. AraNet is a co-functional network derived by integrating 24 -omics datasets from multiple organisms in a machine-learning framework. Both networks showed a similar gradual decrease in fraction of TFs with similar targets with higher JC values (Fig. 3.8A), similar to trends observed in networks with a 'functional context' above. However, we used these networks for comparison only as a rough guide as the AraNet was not designed to prioritize regulatory interactions and holds only approximately 60,000 such edges, and the AtRegNet harbors very few TFs when compared to those in our list. We assumed that both these limitations would make the analysis suffer from the extreme loss of transcriptional signal. However, the robustness of gene relationships predicted in the AraNet was clearly evident as more than 20% of the original TFs in the network presumably interacted even in the highest JC bin, larger than any other networks compared. Overall, the number of TFs observed at any given JC bin in all networks was significantly larger than in a random network. All TF-pairs with JC > 0.70 (arbitrarily chosen stringency) from the seed-specific network were connected and visualized as a graph in Cytoscape (Shannon et al., 2003) revealing many connections supported by multiple networks (Supplemental Fig. S3.6)

About 59% of all genes (23% of all modules) in TMN have at least one known plant CRE enriched in their coexpression neighborhood, with a few modules harboring a large number of different CREs (e.g., Photosynthesis module described earlier) (Fig. 3.8B). Approximately 45%

of total edges in ASCN have an absolute PC score more than 0.9, indicating a highly cohesive network structured for a subtle developmental program.

For evaluation of 'hubs', we selected top 10 TF predictions for each active module in SANe (based on ranked association scores), and counted the number of modules associated with each TF. We observed that 41% of these TFs (552 out of 1339), likely regulate expression of genes in only one module each, while a single TF, NAP57 (AT3G57150), was predicted to be associated with the maximum number of modules (9 out of 120) (Fig. 3.8C). The *NAP57* gene encodes the Arabidopsis dyskerin homolog involved in maintaining telomerase activity (Kannan et al., 2008). As expected, 5 out of 9 modules containing genes whose expression is predicted to be regulated by NAP57 are enriched in GO BP terms such as "DNA metabolic process", "ribonucleoprotein complex biogenesis", "RNA processing" and "ribosome biogenesis". This association was true even on the level of individual targets predictions for majority of the other seed-hubs, in both, the seed and global networks (Table 3.1), indicating that these TFs are responsible for perpetual regulation of important basic processes like biogenesis of cell components, maintenance of cell shape and structure, nucleic acid metabolism etc. A weak but significant enrichment was found between WRKY13 (AT4G39410), a biotic and abiotic stress regulator (Qiu et al., 2007; Xiao et al., 2013), and the GO term 'immune system response' only in the seed network.

Table 3.1: Regulatory hubs of seed development. 23 regulators (transcription factors) that were found associated with the largest number of coexpressed modules in SANe were selected and listed in descending order according to the number of modules they regulate. Targets of these regulators in the seed and the global network, with absolute Z score > 3 were selected and tested for overlaps with BP terms in the GO database. The score columns represent (-1) * log (*q*-value) values from a cumulative hypergeometric test of enrichment. Only the most highly scored gene sets are reported in the table.

Network	**Seed**		**Global**	
Transcription Factor	**Biological Process**	**Enrichment Score**	**Biological Process**	**Enrichment Score**
NAP57 AT3G57150	ribonucleoprotein complex biogenesis	57.05	ribosome biogenesis	72.07
HDT3 (AT5G03740)	ribonucleoprotein complex biogenesis	45.50	ribonucleoprotein complex biogenesis	71.39
AT4G37130	ribonucleoprotein complex biogenesis	40.00	RNA metabolism	35.36
EMB2746 (AT5G63420)	ribonucleoprotein complex biogenesis	56.44	RNA metabolism	29.30
C3H (AT5G60820)	ribonucleoprotein complex biogenesis	12.97	vesicle-mediated transport	10.38
JMJ22 (AT5G06550)	ribonucleoprotein complex biogenesis	36.76	ribosome biogenesis	75.24
WRKY13 (AT4G39410)	immune system process	3.04	N.D	NA
TFIIIA (AT1G72050)	ribosome biogenesis	49.90	RNA metabolism	49.30
VOZ1 (AT1G28520)	ribosome biogenesis	13.30	cellular biopolymer catabolism	4.77
NFD1 (AT4G30930)	ribonucleoprotein complex biogenesis	71.39	ribosome biogenesis	75.24
KAN3 (AT4G17695)	jasmonic acid biosynthesis	4.64	response to salicylic acid stimulus	2.95

Table 3.1 (Cont.)

Network	Seed		Global	
Transcription Factor	**Biological Process**	**Enrichment Score**	**Biological Process**	**Enrichment Score**
HDT1 (AT3G44750)	ribonucleoprotein complex biogenesis	41.25	ribosome biogenesis	72.44
HAT3.1 (AT3G19510)	RNA metabolism	7.88	RNA metabolism	20.48
FZF (AT2G24500)	RNA metabolism	23.67	ribosome biogenesis	68.91
IAA8 (AT2G22670)	polysaccharide metabolism	5.33	transmembrane receptor protein tyrosine kinase signaling pathway	10.47
SMAD/FHA (AT2G21530)	photosynthesis	54.25	photosynthesis	68.43
AT1G78280	cellular biopolymer metabolism	5.88	ribosome biogenesis	8.46
ZFP4 (AT1G66140)	N.D.	N.A.	ion transport	4.40
TRB1 (AT1G49950)	maintenance of root meristem identity	2.33	protein modification	6.33
SEUSS (AT1G43850)	microtubule-based process	2.71	negative regulation of gene expression	8.25
NAC017 (AT1G34190)	vesicle-mediated transport	4.70	vesicle-mediated transport	11.55
ATU2AF35A (AT1G27650)	RNA metabolism	20.60	RNA metabolism	15.07
AT1G17520	proteolysis	3.39	RNA metabolism	12.81

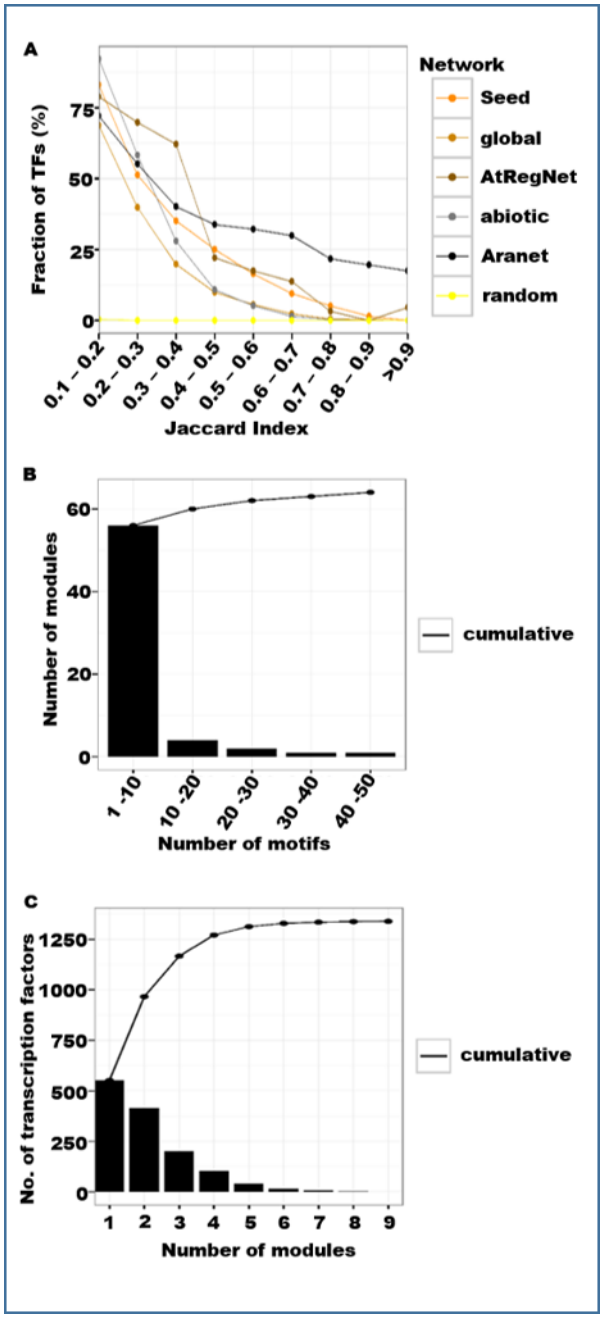

Figure 3.8: Characteristics of seed networks. A) Comparison of the fraction of TFs possibly coregulating the same sets of genes, evaluated using the Jaccard's Index (JI) of overlap between the predicted targets of each TF-pair, for 5 different regulatory networks and a random network. B) A histogram showing bins of number of motifs significantly over-represented in the promoters of genes within each module in TF-MAN. C) Distribution of TF-module edges in SANe follows a scale-free topology, with a large number of regulators associated with fewer modules, and a few regulators (e.g., NAP57, KAN3) associated with a large number of modules.

The SANe webserver

The data generated in this study are represented on a web-based interactive platform available at https://plantstress-pereira.uark.edu/SANe/. The platform allows users to investigate seed development in three different modes (Fig. 3.9):1) Select modules with high expression in compartment – or stage-specific manner, 2) Using the 'cluster enrichment tool' to upload a differential expression profile (e.g. transcriptome of a TF mutant) and identify clusters that significantly perturb in their experiment and 3) enter the locus ID of a TF of interest to identify clusters that are likely regulated by that TF, enabling the user to gain a insight on its functional role prior to an *in vivo* validation. Furthermore, the webserver allows users to visualize the expression of resulting modules/clusters as publication-ready downloadable heatmaps, as well as plot gene connection graphs using Cytoscape (Lopes et al., 2010).

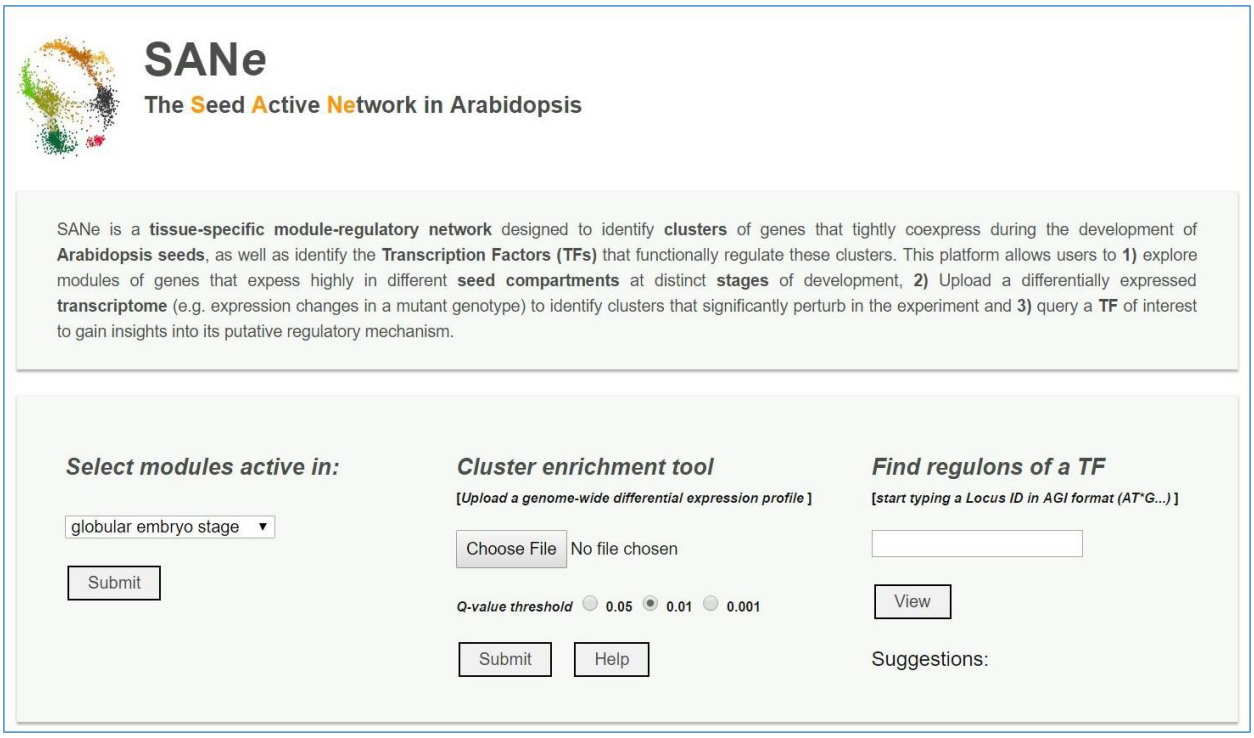

Figure 3.9: Screenshot of the SANe user interface. The SANe web platform (https://plantstress-pereira.uark.edu/SANe/) allows users to identify modules active in distinct seed compartments in different stages of development, upload a new transcriptome in the cluster enrichment tool that uses the parametric analysis to identify enriched modules, and find the regulons of a TF of interest.

Discussion

Plant seeds are complex structures and seed formation is perhaps the most important developmental phase of a plant life cycle, as it determines the fate of the next progeny. Distinct cell types and organs within a seed gradually develop during a period of 20-21 days after pollination in Arabidopsis. In addition, each organ is subjected to its own developmental program and has different, but equally important functions, from feeding and providing optimal growth conditions to protecting the embryo to ensure species propagation. These processes are tightly regulated by synergistically acting TFs (To et al., 2006).

We devised a new methodology that relies on existing statistical methods that are widely accepted, for the discovery of a modular regulatory network. Using a seed-tissue specific expression dataset, this method facilitated identification of modules of coregulated genes, the corresponding development phases in which the modules express most, CREs that drive the biological functions encoded by the genes within modules, and TF regulators that likely govern the expression of the genes in the modules. Our method is limited to making functional predictions for TFs in a tissue-specific manner, and might not accurately predict individual targets of a given TF. This limitation is partly due to the use of a single data-type; a heterogenous approach should be undertaken (e.g. high-throughput DNA binding essays in conjunction with expression data) for studies aiming at specific individual targets. Nevertheless, the statistically significant functional associations predicted here are of superior quality, as seen in evidence from the literature, and can serve as the first step in selecting TFs for targeted downstream experiments. The network inference pipeline presented here can be used to enhance any coexpression based study.

Previous studies have reported a few seed-specific genes, including TFs (Le et al., 2010; Belmonte et al., 2013). We prioritized these genes in our network to derive an active subnetwork,

referred to as Seed Active Network (SANe). We described selected modules containing genes with high expression in specific seed components, including embryo, endosperm and seed coat. We observed that, in most of the cases, the top predicted regulators of these modules are already known in the literature for their involvement in seed development, self-validating our approach. Several additional regulators are known to modulate other processes, including flower development, indicating conserved regulons of pre-fertilization events. Our results suggest that associating regulators to gene sets with a shared function, as opposed to individual genes, provides biologically plausible predictions that are worth for validating *in planta* phenotypes using reverse genetics. As a community resource, our network is accessible through an online platform supported with query driven tools to enable a network based discovery of seed regulatory mechanisms.

It appears that during seed development, photosynthesis and storage compound synthesis is tightly coordinated by several regulators acting coordinately. This was evident from CRE enrichment analysis, as two complementary methods detected the module annotated for photosynthesis and related processes (M0011) harboring genes with the largest number of known plant motifs in their promoters when compared to the rest of the modules. Coordinate regulation of photosynthetic carbon metabolism has been shown previously (Bailey et al., 2007; Ambavaram et al., 2014). Our analysis reveals that much of the processes related to embryo development are conserved throughout the plant life cycle such as cell division and differentiation, as observed by similar roles of regulatory genes in developing embryos and roots. However, plants have developed intrinsic mechanisms that can modulate gene activity in specialized cells, perhaps as duplicated genes with similar functional roles. Such a phenomenon was evident in the case of two TF genes, *NTM1* and *NTM2* that are in close proximity to each other and possibly have similar biological roles in distinct parts of a plant.

The data generated by our work has the potential to further our knowledge of fundamental processes that regulate diverse specific aspects of seed development in Arabidopsis and can be extrapolated to related agriculturally important crops due to conservation of these basic processes (Magallón and Sanderson, 2002; Comparot-Moss and Denyer, 2009; Vriet et al., 2010). Based on our results, a cell- and developmental stage-specific network inference provides superior quality of predictions in the context of known information. Our network analysis pipeline can be further used to systematically increase this information-base for a variety of plant organs (e.g., parts from a post-germination stage network). Comparisons of different stage/tissue specific networks will throw light on the changing molecular mechanisms of a cell and reveal differentially modulated transcriptional networks during different growth stages.

Materials and Methods

Gene expression quantification

Affymetrix ATH1 Arabidopsis gene expression data was downloaded from GEO, and 6 datasets were selected from the super series labeled GSE12404 for seed expression compendium. In addition, 140 other datasets were used in the EEC (Supplemental Table S3.7). All datasets were individually processed in R Bioconductor using a custom CDF file for Arabidopsis (Harb et al., 2010). The re-annotated CDF assigns probe-sets to specific genes and increases the accuracy in expression quantification**.** Using Robust Multi-array average algorithm (RMA) (Irizarry et al., 2003), probe level expression values were background corrected, normalized and summarized into gene level expression values. Values from replicate arrays were then averaged and assembled in an integrated expression matrix of genes as rows and samples as columns, with each cell in the

matrix representing log transformed expression value of genes in the corresponding samples. This procedure resulted in two expression matrices: a seed-specific expression matrix and a global expression matrix.

Coexpression network and cluster identification

Pearson's Correlation (PC) were calculated for each gene pair using expression values in both gene expression matrices. PCs were Fisher Z transformed and standardized to a N(0,1) distribution, where a Z-score of a gene-pair represents the number of standard deviations the score lies away from the mean (Huttenhower et al., 2006). The following procedure was applied only to the seed network. Gene pairs with Z scores above 1.96 (PC 0.75) were retained and connected to create a coexpression network with 21,267 genes connected with approximately 7.6 million edges. SPICi, a fast clustering algorithm (Jiang and Singh, 2010), was used to cluster the network at a range of T_d values ranging from 0.1 to 0.90, keeping a minimum cluster size of 3. Each T_d value was evaluated on three criteria: i) total number of clusters yielded and the fraction of original genes retained in those clusters ii) average modularity following the (Newman and Girvan, 2004) algorithm and iii) functional coherence of clusters based on GO BP term annotations. At T_d 0.80, expression values of each gene within each of 1563 clusters were averaged across the same parts of the seed and in different developmental stages, resulting in two expression profiles for each module. Expression values were scaled and plotted as heatmaps in R using the gplots package (https://CRAN.R-project.org/package=gplots).

Functional annotations of coexpression clusters

The TAIR gene association file was downloaded from the plant GSEA website (http://structuralbiology.cau.edu.cn/PlantGSEA/download.php) (Yi et al., 2013). The .gmt files were filtered to remove generic terms that annotate more than 500 genes, and the remaining list of terms in the BP category were used for testing overlaps with clusters. The significance of overlap of a target gene set (e.g. a cluster) with BP terms was calculated using a cumulative hypergeometric test. The p-values obtained were adjusted for false discovery rate and converted to qvalues using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Enrichment scores were reported as (-1) * log (*q*value).

Analysis of known CREs

We used a pattern-based method to search for CREs over-represented in the promoters of co-regulated genes. First, all known plant motifs were identified from PLACE (Higo et al., 1999) and AGRIS databases (Palaniswamy et al., 2006). Subsequently, 1000-bp upstream promoter regions of all Arabidopsis genes were downloaded from TAIR and scanned for occurrence of these motifs using DNA-pattern matching tool (Medina-Rivera et al., 2015), yielding a list of 403 motifs present at least once in the promoters of ~17000 genes. A few of these motifs, perhaps involved in functions common to all the promoters, are ubiquitously present in almost all the genes. To detect a reliable presence-absence signal in the context of our analysis, we removed motifs that were found in more than 50% of all the genes considered in the network. Thus, a list of 341 unique motifs were used for enrichment (overlap) analysis using a hypergeometric test as described above.

Module Regulatory Network analysis

A list of 1921 Arabidopsis TF regulators was curated from the Plant Transcription Factor Database, the AGRIS database and the Database of Arabidopsis Transcription Factors (Guo et al., 2005; Yilmaz et al., 2011; Jin et al., 2014). For every TF-gene pair, a Z score representing specific correlation score was calculated using the CLR algorithm (Faith et al., 2007). The Parametric Analysis of Geneset Enrichment (PAGE) algorithm (Kim and Volsky, 2005) was used to evaluate enrichment of CLR scored targets of each TF within each module. P-values were calculated form Z scores of enrichment and corrected for FDR using the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995).

Global regulatory network and comparison of different networks

A global regulatory network was constructed the same way as the seed-specific network, except that EEC of 140 datasets was used. The Arabidopsis abiotic stress regulatory network was obtained from (Vermeirssen et al., 2014). Information on interactions reported in AtRegNet and AraNet was downloaded from http://arabidopsis.med.ohio-state.edu/downloads.html and http://www.functionalnet.org/aranet/download.html, respectively. Regulatory interactions (edges with at least one node as a regulator from our list) were identified from AraNet. For all three externally downloaded networks described above, and the global and seed-specific networks from this study, Jaccard coefficient (JC) of overlap in the predicted targets of each regulator pair was calculated using a perl script. JC scores were binned and the fraction of regulators retained from the original individual network within each bin was plotted in R. The random network was created by preserving the node degree and randomly reshuffling all the edges of the seed network.

Network data was parsed using the Sleipnir library (Huttenhower et al., 2008), Network Analysis Tools (NeAT) (Brohee et al., 2008) and scripts written in R and perl.

References

Albert R (2005) Scale-free networks in cell biology. Journal of Cell Science **118:** 4947

Allorent G, Osorio S, Ly Vu J, Falconet D, Jouhet J, Kuntz M, Fernie AR, Lerbs-Mache S, Macherel D, Courtois F, Finazzi G (2015) Adjustments of embryonic photosynthetic activity modulate seed fitness in Arabidopsis thaliana. New Phytologist **205:** 707-719

Ambavaram MM, Krishnan A, Trijatmiko KR, Pereira A (2011) Coordinated activation of cellulose and repression of lignin biosynthesis pathways in rice. *In* Plant Physiol, Vol 155, United States, pp 916-931

Ambavaram MMR, Basu S, Krishnan A, Ramegowda V, Batlang U, Rahman L, Baisakh N, Pereira A (2014) Coordinated regulation of photosynthesis in rice increases yield and tolerance to environmental stress. Nat Commun **5**

Aoki Y, Okamura Y, Tadaka S, Kinoshita K, Obayashi T (2016) ATTED-II in 2016: A Plant Coexpression Database Towards Lineage-Specific Coexpression. Plant Cell Physiol **57:** e5

Bailey KJ, Gray JE, Walker RP, Leegood RC (2007) Coordinate Regulation of Phosphoenolpyruvate Carboxylase and Phosphoenolpyruvate Carboxykinase by Light and CO2 during C4 Photosynthesis. Plant Physiology **144:** 479-486

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucl Acids Res **35**

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

Baud S, Mendoza MS, To A, Harscoet E, Lepiniec L, Dubreucq B (2007) WRINKLED1 specifies the regulatory action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in Arabidopsis. Plant J **50:** 825-838

Belmonte MF, Kirkbride RC, Stone SL, Pelletier JM, Bui AQ, Yeung EC, Hashimoto M, Fei J, Harada CM, Munoz MD, Le BH, Drews GN, Brady SM, Goldberg RB, Harada JJ (2013) Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. Proceedings of the National Academy of Sciences **110:** E435-E444

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57:** 289-300

Boutilier K, Offringa R, Sharma VK, Kieft H, Ouellet T, Zhang L, Hattori J, Liu C-M, van Lammeren AAM, Miki BLA, Custers JBM, van Lookeren Campagne MM (2002) Ectopic Expression of BABY BOOM Triggers a Conversion from Vegetative to Embryonic Growth. The Plant Cell **14:** 1737-1749

Brohee S, Faust K, Lima-Mendez G, Sand O, Janky R, Vanderstocken G, Deville Y, van Helden J (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. Nucleic Acids Res **36:** W444-451

Castillo-Davis CI, Hartl DL (2003) GeneMerge--post-genomic analysis, data mining, and hypothesis testing. Bioinformatics **19:** 891-892

Chavez Montes RA, Coello G, Gonzalez-Aguilera KL, Marsch-Martinez N, de Folter S, Alvarez-Buylla ER (2014) ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks. *In* BMC Plant Biol, Vol 14, England, p 97

Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C (2011) Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLoS ONE **6:** e17238

Chen L-Q, Lin IW, Qu X-Q, Sosso D, McFarlane HE, Londoño A, Samuels AL, Frommer WB (2015) A Cascade of Sequentially Expressed Sucrose Transporters in the Seed Coat and Endosperm Provides Nutrition for the Arabidopsis Embryo. The Plant Cell **27:** 607-619

Childs KL, Davidson RM, Buell CR (2011) Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. PLoS ONE **6:** e22196

Comparot-Moss S, Denyer K (2009) The evolution of the starch biosynthetic pathway in cereals and other grasses. Journal of Experimental Botany **60:** 2481-2492

Devaiah BN, Madhuvanthi R, Karthikeyan AS, Raghothama KG (2009) Phosphate Starvation Responses and Gibberellic Acid Biosynthesis Are Regulated by the MYB62 Transcription Factor in Arabidopsis. Molecular Plant **2:** 43-58

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences **95:** 14863-14868

Eshed Y, Izhaki A, Baum SF, Floyd SK, Bowman JL (2004) Asymmetric leaf development and blade expansion in Arabidopsis are mediated by KANADI and YABBY activities. Development **131:** 2997-3006

Fait A, Angelovici R, Less H, Ohad I, Urbanczyk-Wochniak E, Fernie AR, Galili G (2006) Arabidopsis Seed Development and Germination Is Associated with Temporally Distinct Metabolic Switches. Plant Physiology **142:** 839-854

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-Scale Mapping and Validation of <named-content xmlns:xlink="http://www.w3.org/1999/xlink" content-type="genus-species" xlink:type="simple">Escherichia coli</named-content> Transcriptional Regulation from a Compendium of Expression Profiles. PLoS Biol **5:** e8

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

Goffman FD, Alonso AP, Schwender J, Shachar-Hill Y, Ohlrogge JB (2005) Light enables a very high efficiency of carbon storage in developing embryos of rapeseed. Plant Physiol **138:** 2269-2279

Goffman FD, Ruckle M, Ohlrogge J, Shachar-Hill Y (2004) Carbon dioxide concentrations are very high in developing oilseeds. Plant Physiol Biochem **42:** 703-708

González-Morales SI, Chávez-Montes RA, Hayano-Kanashiro C, Alejo-Jacuinde G, Rico-Cambron TY, de Folter S, Herrera-Estrella L (2016) Regulatory network analysis reveals novel regulators of seed desiccation tolerance in Arabidopsis thaliana. Proceedings of the National Academy of Sciences **113:** E5232-E5241

Grimault A, Gendrot G, Chaignon S, Gilard F, Tcherkez G, Thévenin J, Dubreucq B, Depège-Fargeix N, Rogowsky PM (2015) Role of B3 domain transcription factors of the AFL family in maize kernel filling. Plant Science **236:** 116-125

Grossniklaus U, Vielle-Calzada JP, Hoeppner MA, Gagliano WB (1998) Maternal control of embryogenesis by MEDEA, a polycomb group gene in Arabidopsis. Science **280:** 446-450

Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J (2005) DATF: a database of Arabidopsis transcription factors. Bioinformatics **21:** 2568-2569

Harb A, Krishnan A, Ambavaram MMR, Pereira A (2010) Molecular and Physiological Analysis of Drought Stress in Arabidopsis Reveals Early Responses Leading to Acclimation in Plant Growth. Plant Physiology **154:** 1254-1271

He L, Su C, Wang Y, Wei Z (2015) ATDOF5.8 protein is the upstream regulator of ANAC069 and is responsive to abiotic stress. Biochimie **110:** 17-24

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

Husbands AY, Benkovics AH, Nogueira FTS, Lodha M, Timmermans MCP (2015) The ASYMMETRIC LEAVES Complex Employs Multiple Modes of Regulation to Affect Adaxial-Abaxial Patterning and Leaf Complexity[OPEN]. The Plant Cell

Huttenhower C, Hibbs M, Myers C, Troyanskaya OG (2006) A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics **22:** 2890-2897

Huttenhower C, Mutungu KT, Indik N, Yang W, Schroeder M, Forman JJ, Troyanskaya OG, Coller HA (2009) Detailing regulatory networks through large scale data integration. Bioinformatics **25:** 3267-3274

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

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE **5:** e12776

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249-264

Izhaki A, Bowman JL (2007) KANADI and Class III HD-Zip Gene Families Regulate Embryo Patterning and Modulate Auxin Flow during Embryogenesis in Arabidopsis. The Plant Cell **19:** 495-508

Jia H, McCarty DR, Suzuki M (2013) Distinct Roles of LAFL Network Genes in Promoting the Embryonic Seedling Fate in the Absence of VAL Repression. Plant Physiology **163:** 1293-1305

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

Jin J, Zhang H, Kong L, Gao G, Luo J (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Research **42:** D1182-D1187

Johnson CS, Kolevski B, Smyth DR (2002) TRANSPARENT TESTA GLABRA2, a Trichome and Seed Coat Development Gene of Arabidopsis, Encodes a WRKY Transcription Factor. The Plant Cell **14:** 1359-1375

Kannan K, Nelson AD, Shippen DE (2008) Dyskerin is a component of the Arabidopsis telomerase RNP required for telomere maintenance. Mol Cell Biol **28:** 2332-2341

Keith K, Kraml M, Dengler NG, McCourt P (1994) fusca3: A Heterochronic Mutation Affecting Late Embryo Development in Arabidopsis. The Plant Cell **6:** 589-600

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

Kim Y-S, Kim S-G, Park J-E, Park H-Y, Lim M-H, Chua N-H, Park C-M (2006) A Membrane-Bound NAC Transcription Factor Regulates Cell Division in Arabidopsis. The Plant Cell **18:** 3132-3144

Kohler C, Hennig L, Bouveret R, Gheyselinck J, Grossniklaus U, Gruissem W (2003) Arabidopsis MSI1 is a component of the MEA/FIE Polycomb group complex and required for seed development. Embo j **22:** 4804-4814

Kwong RW, Bui AQ, Lee H, Kwong LW, Fischer RL, Goldberg RB, Harada JJ (2003) LEAFY COTYLEDON1-LIKE Defines a Class of Regulators Essential for Embryo Development. The Plant Cell **15:** 5-18

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9:** 559

Lashbrooke JG, Cohen H, Levy-Samocha D, Tzfadia O, Panizel I, Zeisler V, Massalha H, Stern A, Trainotti L, Schreiber L, Costa F, Aharoni A (2016) MYB107 and MYB9 Homologs Regulate Suberin Deposition in Angiosperms. The Plant Cell

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, Drews GN, Fischer RL, Okamuro JK, Harada JJ, Goldberg RB (2010) Global analysis of gene activity during Arabidopsis seed development and identification of seed-specific transcription factors. Proceedings of the National Academy of Sciences **107:** 8063-8070

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

Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD (2010) Cytoscape Web: an interactive web-based network browser. Bioinformatics **26**

Lotan T, Ohto M, Yee KM, West MA, Lo R, Kwong RW, Yamagishi K, Fischer RL, Goldberg RB, Harada JJ (1998) Arabidopsis LEAFY COTYLEDON1 is sufficient to induce embryo development in vegetative cells. Cell **93:** 1195-1205

Luerßen H, Kirik V, Herrmann P, Miséra S (1998) FUSCA3encodes a protein with a conserved VP1/ABI3-like B3 domain which is of functional importance for the regulation of seed maturation inArabidopsis thaliana. **15:** 755-764

Luo M, Dennis ES, Berger F, Peacock WJ, Chaudhury A (2005) MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America **102:** 17531-17536

Ma H-W, Buer J, Zeng A-P (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics **5:** 199

Magallón S, Sanderson MJ (2002) Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. American Journal of Botany **89:** 1991-2006

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics **7:** S7-S7

McAbee JM, Hill TA, Skinner DJ, Izhaki A, Hauser BA, Meister RJ, Venugopala Reddy G, Meyerowitz EM, Bowman JL, Gasser CS (2006) ABERRANT TESTA SHAPE encodes a KANADI family member, linking polarity determination to separation and growth of Arabidopsis ovule integuments. **46:** 522-531

Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier-Khamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M, van Helden J (2015) RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Research

Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L (2001) The Arabidopsis TT2 Gene Encodes an R2R3 MYB Domain Protein That Acts as a Key Determinant for Proanthocyanidin Accumulation in Developing Seed. The Plant Cell **13:** 2099-2114

Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Physical Review E **69:** 026113

Nygaard V, Rødland EA, Hovig E (2015) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics

Obayashi T, Kinoshita K (2011) COXPRESdb: a database to compare gene coexpression in seven model animals. Nucleic Acids Research **39:** D1016-D1022

Ogas J, Kaufmann S, Henderson J, Somerville C (1999) PICKLE is a CHD3 chromatin-remodeling factor that regulates the transition from embryonic to vegetative development in Arabidopsis. Proceedings of the National Academy of Sciences **96:** 13839-13844

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 Physiology **140:** 818-829

Portereiko MF, Lloyd A, Steffen JG, Punwani JA, Otsuga D, Drews GN (2006) AGL80 Is Required for Central Cell and Endosperm Development in Arabidopsis. The Plant Cell **18:** 1862-1872

Qiu D, Xiao J, Ding X, Xiong M, Cai M, Cao Y, Li X, Xu C, Wang S (2007) OsWRKY13 Mediates Rice Disease Resistance by Regulating Defense-Related Genes in Salicylate- and Jasmonate-Dependent Signaling. Molecular Plant-Microbe Interactions **20:** 492-499

Ruuska SA, Girke T, Benning C, Ohlrogge JB (2002) Contrapuntal Networks of Gene Expression during Arabidopsis Seed Filling. The Plant Cell **14:** 1191-1206

Ruuska SA, Schwender J, Ohlrogge JB (2004) The Capacity of Green Oilseeds to Utilize Photosynthesis to Drive Biosynthetic Processes. Plant Physiology **136:** 2700-2709

Sato Y, Namiki N, Takehisa H, Kamatsuki K, Minami H, Ikawa H, Ohyanagi H, Sugimoto K, Itoh J-I, Antonio BA, Nagamura Y (2012) RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Research

Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y (2004) Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature **432:** 779-782

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**

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

Sreenivasulu N, Wobus U (2013) Seed-development programs: a systems biology-based comparison between dicots and monocots. Annu Rev Plant Biol **64:** 189-217

Suzuki M, McCarty DR (2008) Functional symmetry of the B3 network controlling seed development. Current Opinion in Plant Biology **11:** 548-553

To A, Valon C, Savino G, Guilleminot J, Devic M, Giraudat J, Parcy F (2006) A Network of Local and Redundant Gene Regulation Governs Arabidopsis Seed Maturation. The Plant Cell **18:** 1642-1651

Vermeirssen V, De Clercq I, Van Parys T, Van Breusegem F, Van de Peer Y (2014) Arabidopsis Ensemble Reverse-Engineered Gene Regulatory Network Discloses Interconnected Transcription Factors in Oxidative Stress. The Plant Cell **26:** 4656-4679

Vriet C, Welham T, Brachmann A, Pike M, Pike J, Perry J, Parniske M, Sato S, Tabata S, Smith AM, Wang TL (2010) A Suite of Lotus japonicus Starch Mutants Reveals Both Conserved and Novel Features of Starch Metabolism. Plant Physiology **154:** 643-655

Wendrich JR, Weijers D (2013) The Arabidopsis embryo as a miniature morphogenesis model. New Phytologist **199:** 14-25

Xiao J, Cheng H, Li X, Xiao J, Xu C, Wang S (2013) Rice WRKY13 Regulates Cross Talk between Abiotic and Biotic Stress Signaling Pathways by Selective Binding to Different cis-Elements. Plant Physiology **163:** 1868-1882

Yamamoto A, Kagaya Y, Toyoshima R, Kagaya M, Takeda S, Hattori T (2009) Arabidopsis NF-YB subunits LEC1 and LEC1-LIKE activate transcription by interacting with seed-specific ABRE-binding factors. **58:** 843-856

Yeung KY, Haynor DR, Ruzzo WL (2001) Validating clustering for gene expression data. Bioinformatics **17:** 309-318

Yi X, Du Z, Su Z (2013) PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res **41:** W98-103

Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E (2011) AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. Nucleic Acids Res **39:** D1118-1122

Yim WC, Yu Y, Song K, Jang CS, Lee B-M (2013) PLANEX: the plant co-expression database. BMC Plant Biology **13:** 83

Yu X, Li L, Zola J, Aluru M, Ye H, Foudree A, Guo H, Anderson S, Aluru S, Liu P, Rodermel S, Yin Y (2011) A brassinosteroid transcriptional network revealed by genome-wide identification of BESI target genes in Arabidopsis thaliana. Plant J **65:** 634-646

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR. Arabidopsis Microarray Database and Analysis Toolbox. Plant Physiology **136:** 2621-2632

Chapter 4: Edge-set Enrichment Analysis based on Density: A New Paradigm for Identification of Genes and Pathways from Differential Coexpression.

Abstract

Analysis of differential expression from high-throughput expression data coupled with Gene Set Enrichment Analysis (GSEA) is a widely used metric to evaluate how biologically coherent sets of genes respond to a treatment, leading to identification of pathways and cellular processes directly affected. However, considering every gene individually and disregarding functional relationships between genes in traditional GSEA analysis methods underplay several truly associated biological processes. This study devised a new network-based enrichment analysis strategy that considers functional relationships between pairs of genes to score functionally coherent gene sets. The algorithm uses differential coexpression of gene pairs instead of differential expression of individual genes, and leverages on the change in network density for estimation of gain or loss in correlation of pathways and biological processes. The approach was tested by analyzing expression datasets pertaining specifically to drought or water deficit response in the crop model *Oryza sativa*. Comparing the drought network with an independent control network revealed several emergent properties of the rice drought responses, as well as biological processes and metabolic pathways that remained elusive in traditional differential expression analysis of singular datasets.

Introduction

Expression profiling of genes is one of most widely used molecular assay in modern genomics research. Besides providing information about the genes that are, for example, affected by stress or change in developmental stages, expression data also reveals key cellular pathways that are involved. It is increasingly evident that genes work as groups in pathways, and a change in expression of single individuals in the group might not have a significant perturbation on the associated pathway for an observable phenotype. Therefore, analysis of different gene sets (sets of genes with a unifying biological theme) is crucial for a holistic understanding of how different components of the cell organize and respond to a stimuli. Usually, for well-studied organisms, curated gene sets are available from computationally tractable catalogs such as the Gene Ontology (GO) (Harris et al., 2004) or metabolic pathway databases (Kanehisa et al., 2004; Caspi et al., 2012). The overlap between these gene sets and the list of differentially expressed genes is computed (gene set enrichment analysis) to find pathways that contain a large proportion of perturbed genes. The gene sets can also be customized, for example, to expand on the sparsely annotated GO terms of the genome of interest (Krishnan et al., 2017).

Although analysis of Differential Expression (DE) has been very successful in identifying genes that respond to the experimental setup, not all genes that are responsible for the "trait" of interest are readily available through analysis of DE alone. An observed phenotype is the product of intricate interactions between specific genes (functional relationships that can be direct biophysical interactions or indirect interactions involving one or more intermediate genes), which is the fundamental principle of metabolic pathways and core enzymatic reactions in a biological cell. A subtle change in expression of a single gene can lead to altered relationships with its functionally interacting partners, which can have a profound effect on the ultimate output of the

underlying pathway(s) they together participate in. For example, in signal transduction cascades, small changes in the expression of a few genes – especially the ones with regulatory roles that participate higher up in the hierarchy of the gene network – might result in an amplified signal received by genes that process the signal further downstream in the hierarchy (Klebanov et al., 2006). Hence, it is important to consider functional relationships between genes, and how these relationships change when subjected to change in biological conditions.

A genome-wide coexpression network analysis is a popular approach conducive to estimating the probability of a gene pair being functionally related (D'haeseleer et al., 2000; Stuart et al., 2003; Childs et al., 2011; Yang et al., 2014; Krishnan et al., 2017). However, gene coexpression networks are usually 'static', in the sense that they are either assembled from any usable expression data that is available for the organism of choice (Atias et al., 2009; Childs et al., 2011; Liang et al., 2014), or for a specific 'context' like a single tissue (Rosa et al., 2014; Pierson et al., 2015; Gupta et al., 2017) or environmental condition (Bassel et al., 2011; Shaik and Ramakrishna, 2013; Sircar and Parekh, 2015; Li et al., 2016; Krishnan et al., 2017). However, interactions between genes are 'promiscuous' and change dramatically between different biological contexts (Li, 2002), leading to changes in their relationships with other genes. For example, it is possible that a relationship between a gene pair is conditional, and if the data for this condition is not equally represented in the integrated dataset, the edge might be missed in statistical calculations. Nevertheless, with the right sampling of expression datasets from the ones being accumulated in public resources (Barrett et al., 2007), it becomes somewhat intuitive to compare two networks, for example, 'control' and 'condition' coexpression networks, and consider the differential coexpression of gene pairs, instead of estimating change in individual genes.

While DE examines the change of gene expression between experimental groups, 'Differential Coexpression' (DC) investigates how coexpression between genes change in different conditions. Several statistical models have been proposed to evaluate differences in coexpression patterns between gene pairs (Lai et al., 2004; Lui et al., 2015; Gao et al., 2016; Liany et al., 2017) or differences in a functional module/cluster based inference setting (Chia and Karuturi, 2010; Tesson et al., 2010). Some of these methods have been successfully employed to gain insights into human diseases (de la Fuente, 2010; Walley et al., 2012; Gaiteri et al., 2014; Xu et al., 2015; Yuan et al., 2015), as powerful and useful approaches to complement traditional DE analysis. In plants, differential coexpression has been used to decipher defense response in biotic and abiotic-stresses in Arabidopsis (Ma et al., 2014; Jiang et al., 2016) and tissue-specificity in tomato (Fukushima et al., 2012).

For identification of pathways from DC data, one proposed solution is the Edge Set Enrichment Analysis (ESEA) algorithm that considers a pathway structure and differential correlation among corresponding annotated genes (Han et al., 2015). The ESEA algorithm ranks edges (instead of genes) on the basis of change in correlation to evaluate the top ranked pathways and uses the Kolmogrov-Smirnov statistic to score for statistical significance. Several other edge-centric methods have been proposed for identification of dysregulated pathways (Choi and Kendziorski, 2009; Zhang et al., 2009; Liu et al., 2012). Most of these computational methods use permutations and random shuffling of expression matrices to estimate the statistical significance of observed enrichments. Actually, the number of permutations dictate the stringency of the statistical test as the lowest *p*-value cannot be lesser than *1/n*, where *n* is the total number of permutations.

Herein, a new strategy for identification of genes and pathways from coexpression data is proposed and tested in the crop model *Oryza sativa* (rice). First, the differences in coexpression relationships between gene-pairs was estimated using two conditionally independent coexpression networks. Edges (coexpressed gene pairs) that significantly changed between a control network and a drought network were extracted and genes that frequently participated in these rewired edges were identified as differentially coexpressed genes (DCG). Since the networks were created independent of each other, the current state of functional annotations in rice were leveraged upon, and the coexpression scores within each network were summarized to 'function' level association scores using two methods, each of which represented the strength of coexpression amongst genes co-annotated within a functional domain. Functional domains were cumulatively derived from annotations provided by the Biological Process (BP) category of the Gene Ontology (GO) consortium (Lewis, 2005), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database, together representing a broad spectrum of largely all the biological process and metabolic pathway level classes that were computationally tractable. The workflow is depicted in Figure 4.1.

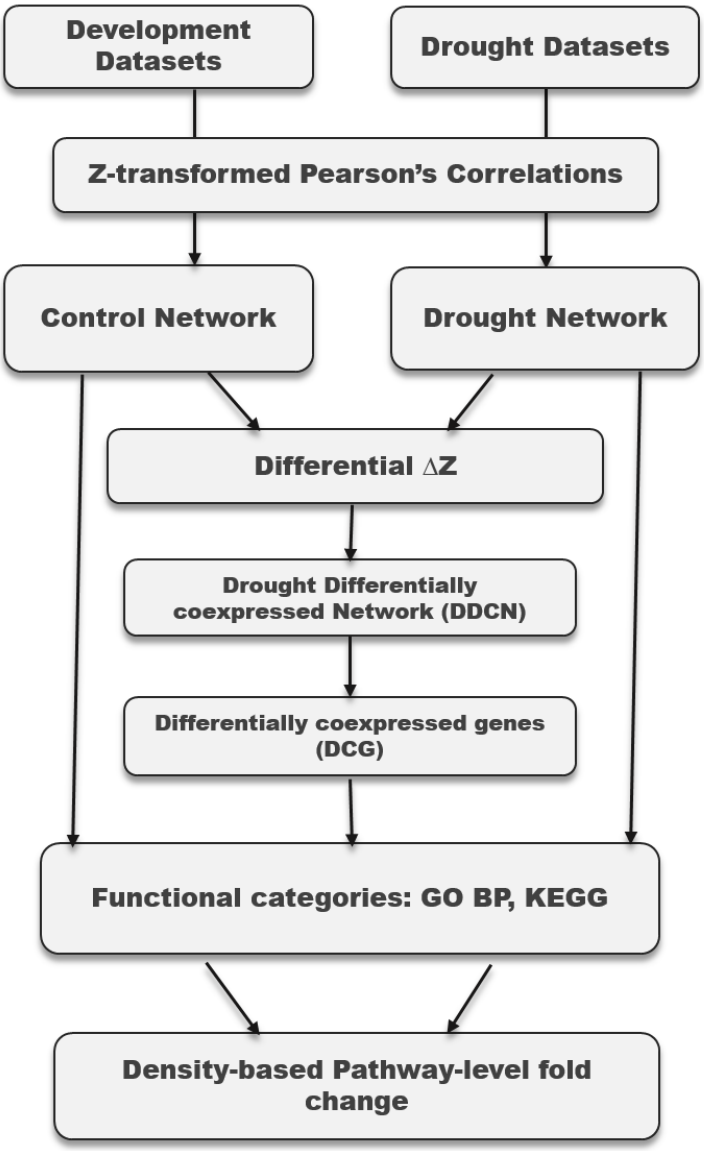

Figure 4.1: The workflow for mining differentially coexpressed genes and pathways. The pipeline used to derive a differentially coexpressed network from two non-redundant rice gene expression compendia. From these, two conditionally independent coexpression networks were assembled and a differential of coexpression scores (ΔZ) between the two networks was taken to derive the Drought Differentially Coexpressed Network (DDCN). The DDCN was examined for Differentially Coexpressed Genes (DCG) that had significantly large number of connections. Network density based measures were used to find the fold change of interaction between and within pathways in both the networks.

Results

Comparison of networks

Two independent coexpression networks were created from rice expression datasets available in the GEO. A drought network (DN) was assembled from samples pertaining to water stress conditions and the control network (CN) was assembled from samples with "non-stressed" labeled experiments (expression quantified in different tissues at various developmental stages). Normalized Pearson's Correlation scores (Huttenhower et al., 2006) were used to compute the two coexpression profiles for each gene-pair. The resulting *Z*-score is the standard score, expressed as the number of standard deviations a coexpressed gene pair lies away from the mean of all scores of that dataset/network, and efficiently handles possible batch effects in the data (Cheadle et al., 2003).

A gene-pair was classified as differentially coexpressed if there was a significant difference in their coexpression scores between the DN and the CN. The difference in pair-wise correlation scores was computed as a ΔZ score (see equation 2 in "Methods"). Since the ΔZ distribution will always follow a Gaussian distribution (Fukushima, 2013), *p*-values were calculated under the standard normal distribution, and ~3.1 million rewired edges with a stringent cutoff ($|Z| > 5$) were extracted. These edges are referred to as the Drought Differentially Coexpressed Network (DDCN) for the remainder of this paper.

Differentially Coexpressed Genes (DCG) were then defined as those that frequently participated (enriched) at the edges of the DDCN. Specifically, first the number of connections per gene in the DDCN were computed. Then, for a gene *i* that participated in *k* differentially coexpressed gene-pairs, the probability of observing that gene *k* times was estimated using a

binomial probability model (Jiang et al., 2016). A total of 15163 genes were identified as DCG (FDR corrected *p*-value < 1×10^{-7}) from the DDCN. These DCG, sorted by the number of connections, are listed in Supplemental table S4.1.

The next question was identification of cellular pathways represented in the DDCN and the DCG. Since the study presented here dealt with two large networks, permutation of the two expression matrices, if not impossible, would be computationally expensive and extremely time-consuming. Instead, to investigate drought mediated alterations on rice metabolism, the focus was set on evaluating and modifying different procedures and techniques traditionally used in enrichment analysis.

Gene sets enriched in DCG reveal much more then DE genes

The first question was whether the identified DCG were similar to the genes that could be detected on the basis of Differential Expression (DE). To evaluate the commonalities, rice drought samples of three development stages from the dataset GSE81253 were analyzed to identify genes that significantly DE. The overlaps between the DCGs and the DE genes at all the stages were then visualized using an UpSet plot (Conway et al., 2017). It was observed that the number of DCGs were significantly large in number (p <0.001) as compared to the number of genes with significant DE in any of the three developmental stages. The largest overlap between DCGs and a developmental stage was at the seedling stage with ~30% genes (3609) in common (Fig 4.2 A). A total of 4519 genes were identified as unique in the list of DCG, with only 726 genes common to all four lists. Even at the level of functional information, DCGs harbored the highest number of unique GO BP terms (41) as compared to any of the three individual lists of DE genes, as evaluated

using a Fisher's exact test. Again, the largest overlap of BPs enriched in a DE gene list with DCG was at the seedling stage with 24 BPs commonly enriched, while 78 BPs were found common to all four lists (Fig. 4.2 B).

Some of the BPs that were unique to the list of DCGs are well known drought responses, but could not be revealed from the list of DE genes at any of the three developmental stages tested. For example, the GO term "aromatic compound biosynthesis" was found enriched only in the list of DCGs (*p*-value 0.0366). The up-regulation of aromatic amino acids tryptophan, tyrosine and phenylalanine were found increased in maize and wheat leaves under drought stress (Bowne et al., 2012; Witt et al., 2012). The enrichment of the term "iron-sulfur cluster assembly" (*p*-value 0.0356) in DCGs is also attributed to drought stress as sulfur use efficiency has been previously linked with drought in *Brassica napus* (Lee et al., 2016). Some other known mechanisms, like "regulation of GTPase activity" (*p*-value 4.9×10^{-3}), related to signal transduction under drought (Ferrero-Serrano and Assmann, 2016), could also be only revealed from enrichment analysis of DCG genes. The list of BPs terms unique to the list of DCGs are listed in table 4.1. Altogether, this indicated that differential coexpression indeed detected a larger number of genes and BPs that were responsive to drought, and could not be detected from DE analysis.

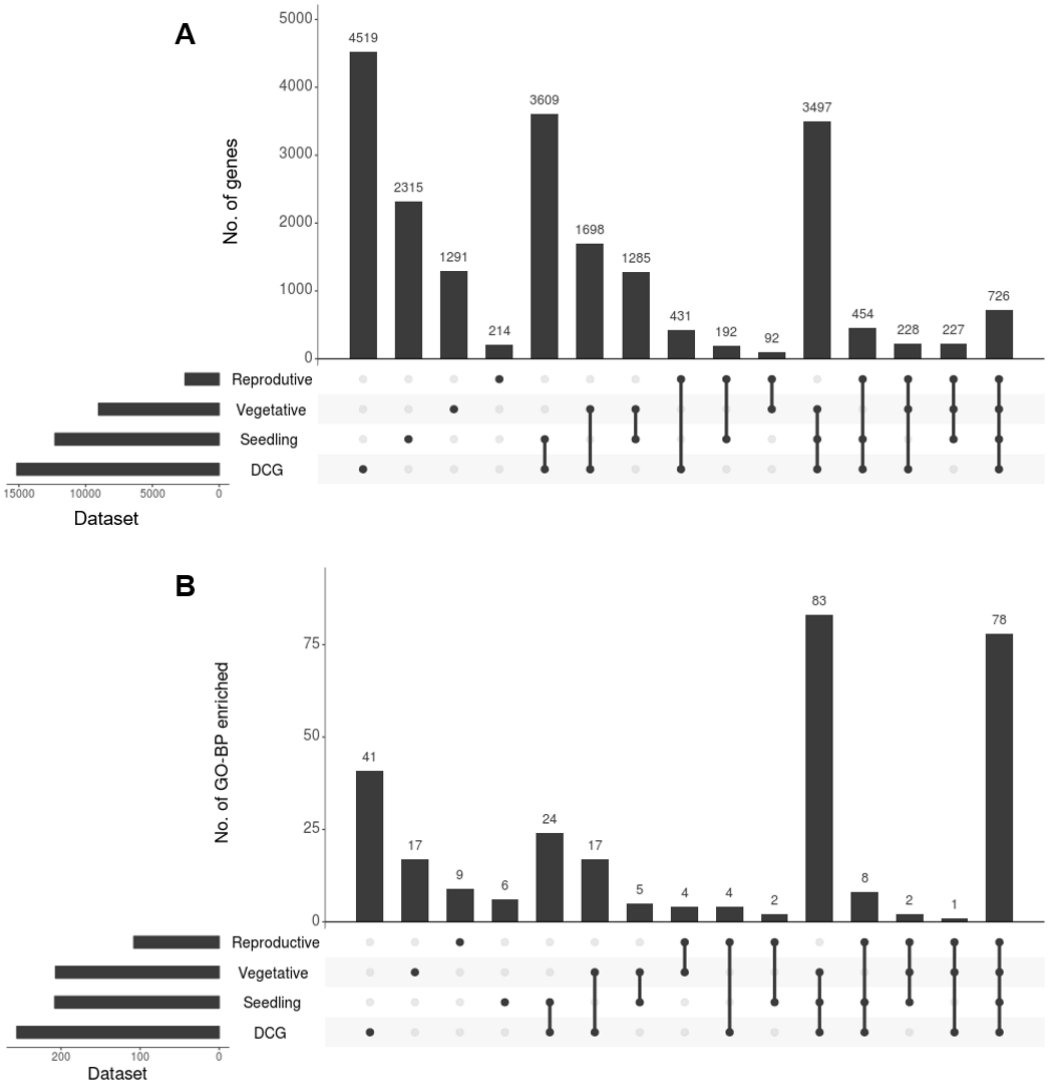

Figure 4.2: UpSet plots illustrating the overlaps between differential coexpression and differential expression analysis. A) Differentially coexpressed genes were identified from the DDCN and overlapped with differentially expressed genes from three stages of drought treatments: Reproductive, Vegetative and Seedling. B) The overlaps between significantly enriched GO BP terms in the list of DCGs and DE genes in all the three stages. The UpSet plots served as an alternative to Venn Diagrams for comparision of overlaps between a large number of sets. In both the plots depicted here, the bars on the lower left correspond to the total number of genes or BPs within the indicated sample name. The bars on the top indicate the intersection between sets. The sets being compared are illustrated by the matrix on the bottom where individual dots against the sample name indicate unique items in the sample and lines connecting the dots indicate the overlap between the corresponding sets, with the magnitude of the overlap indicated in the top bar plot. The number above each bar plot indicate the number of overlapping items.

Table 4.1: Biological processes that cannot be detected from the lists of differentially expressed genes but found enriched in the list of genes that are differentially coexpressed. A table of biological processes (first column) that were identified as unique in the list of differentially coexpressed genes in the DDCN, along with the false discovery rate corrected p-values obtained from the Fisher's exact test (second column).

GO BP Gene set	***p* value**
mRNA metabolic process	9.92E-05
glutamine family amino acid metabolic process	6.93E-04
phospholipid metabolic process	8.78E-04
protein targeting	1.61E-03
cell wall organization	2.50E-03
hydrogen transport	3.71E-03
proton transport	3.71E-03
regulation of GTPase activity	4.93E-03
cellular protein complex assembly	7.82E-03
chromatin assembly	0.0101
nucleosome organization	0.0101
protein-DNA complex assembly	0.0101
Co-translational protein targeting to membrane	0.0104
SRP-dependent co-translational protein targeting to membrane	0.0104
cytoskeleton organization	0.0152
protein polymerization	0.02
RAS protein signal transduction	0.0213
regulation of small GTPase mediated signal transduction	0.0213
regulation of cell communication	0.024
regulation of signal transduction	0.024
cellular membrane organization	0.028
membrane organization	0.028
iron-sulfur cluster assembly	0.0356
Metallo-sulfur cluster assembly	0.0356
aromatic compound biosynthetic process	0.0366
nucleoside triphosphate metabolic process	0.0454

Edge-set enrichment analysis using network density

The density of a network reflects the cohesiveness between the underlying genes in the network, and is measured as the number of edges observed in a network divided by the total number of expected edges. Within a range of 0 and 1, density values near 1 indicates a fully connected network (majority of the genes interacting with each other) reflecting more coherence between genes as compared to a network with low density. In the context of differential coexpression analysis here, an interesting question was whether there were differences in the densities, meaning whether there is a gain of correlation or loss of correlation between the DN and the CN. Therefore, to evaluate how the loss of edges in the DN translated to functional pathways, the weighted densities of subnetworks induced by genes annotated in each of the rice KEGG pathways and GO BPs were vetted in both the networks.

Using this concept of network density, a new formulation was derived to score gene set coherence in an underlying network. The algorithm is referred to as Edge-set Enrichment Analysis based on Density (EsEAD; pronounced 'assayed'). EsEAD starts by calculating a weighted density of each of the given gene sets within a network, where the resulting density scores indicated the overall coherence of the pathway as a whole. Next, EsEAD proceeds with estimating the fold change in empirically calculated densities between the two input networks (see Methods). Under this formulation, it appeared that genes functioning in the same functional class are expressed in a less cohesive manner in the DN as the average intra-pathway density (edges co-annotated to the same class) dropped slightly as compared to the CN, in both the annotation catalogs (Fig. 4.3 A and B). However, the differences in the median density of all classes was statistically insignificant. On the other hand, the average inter-pathway cohesiveness (edges with genes cross-annotated to different functional domains) was found to be higher in the DN as compared to the CN (Fig. 4.3

C and D). On closer examination of the pathway gene sets from KEGG, it was observed that 307 new inter-pathway connections that were absent in the CN appeared in the DN. Some of these inter-pathway connections were obvious drought responses; the appearance of communication between the "Cutin, suberin and wax biosynthesis" pathway and "fatty acid metabolism" is suggestive of a drought resistance mechanism by increased wax biosynthesis to maintain water loss (Chen et al., 2011). Cytokinin mediated alteration of source/sink relationship under drought was also observed as the "Zeatin biosynthesis" pathway and "starch and sucrose metabolism" communicated only in the DN, validating previous drought response observation (Peleg et al., 2011). Other such drought-specific and common inter-pathway connections, sorted by density, are listed in the Supplemental table S4.2 and Supplemental S4.3, respectively.

In terms of intra-pathway cohesiveness, majorly all pathways had a decrease in the fold change of density in the DN (Fig. 4.4). The "Porphyrin and chlorophyll metabolism" and the "plant pathogen interaction" pathways showed almost 2-fold decrease in the density, followed by the "Glyoxylate and dicarboxylate metabolism" pathway and other pathways related to photosynthesis. In contrast, amongst the few pathways that were up-regulated, the "N-Glycan biosynthesis" pathway had the highest fold change increase in density (1.5 fold) as compared to CN. Glycans have been implicated under drought in a variety of plants under drought stress (Fracasso et al., 2016; Muthusamy et al., 2016). The "Valine, leucine and isoleucine degradation" pathway was also found upregulated in the DN, supported by the increased levels of branched chain amino acids in response to drought stress that has been previously detected in other cereals like wheat (Bowne et al., 2012). The full list of intra-pathway densities, sorted by fold change values, are listed in Supplemental table S4.4.

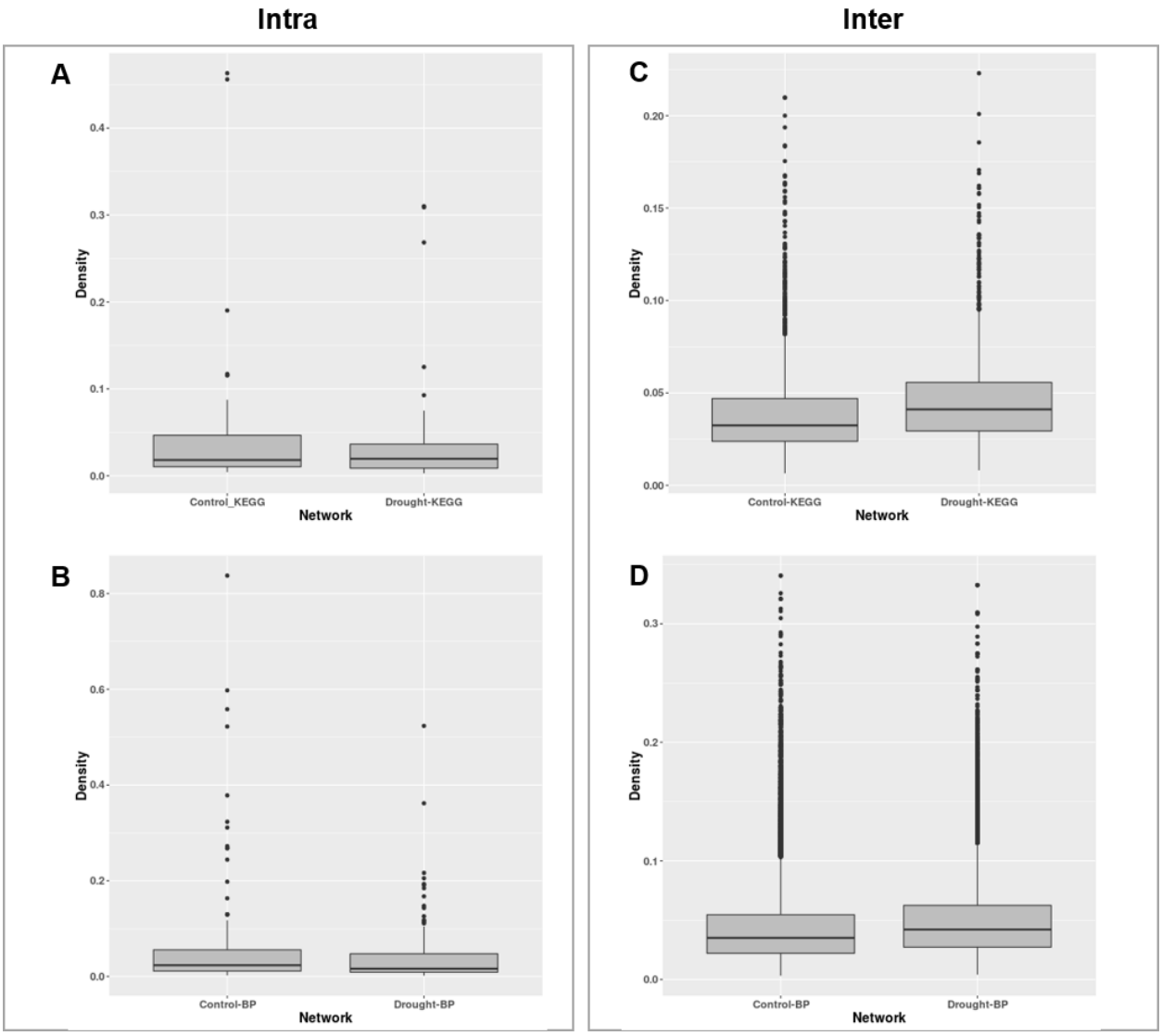

Figure 4.3: Comparison of the network densities within the two functional annotation catalogs of rice considered in this study. The coherence between genes within a functional class was evaluated using density, where a higher density indicates more cohesiveness between genes of a pathway. The overall densities were plotted as boxplots and the center of the box corresponds to the median (2nd quartile; Q_2) of the distribution of densities (Y axis). The extremes of the box correspond to the 1st (Q_1) and 3rd (Q_3) quartiles. The whiskers denote $Q_2 \pm 1.5*IQR$, where IQR is the interquartile range (Q_3-Q_1). The left panel shows the intra-pathway densities of A) 117 KEGG annotations and B) 634 GO BP terms, in both the networks compared. The right panel show the inter-pathway density of the same C) KEGG annotations and D) GO BP terms.

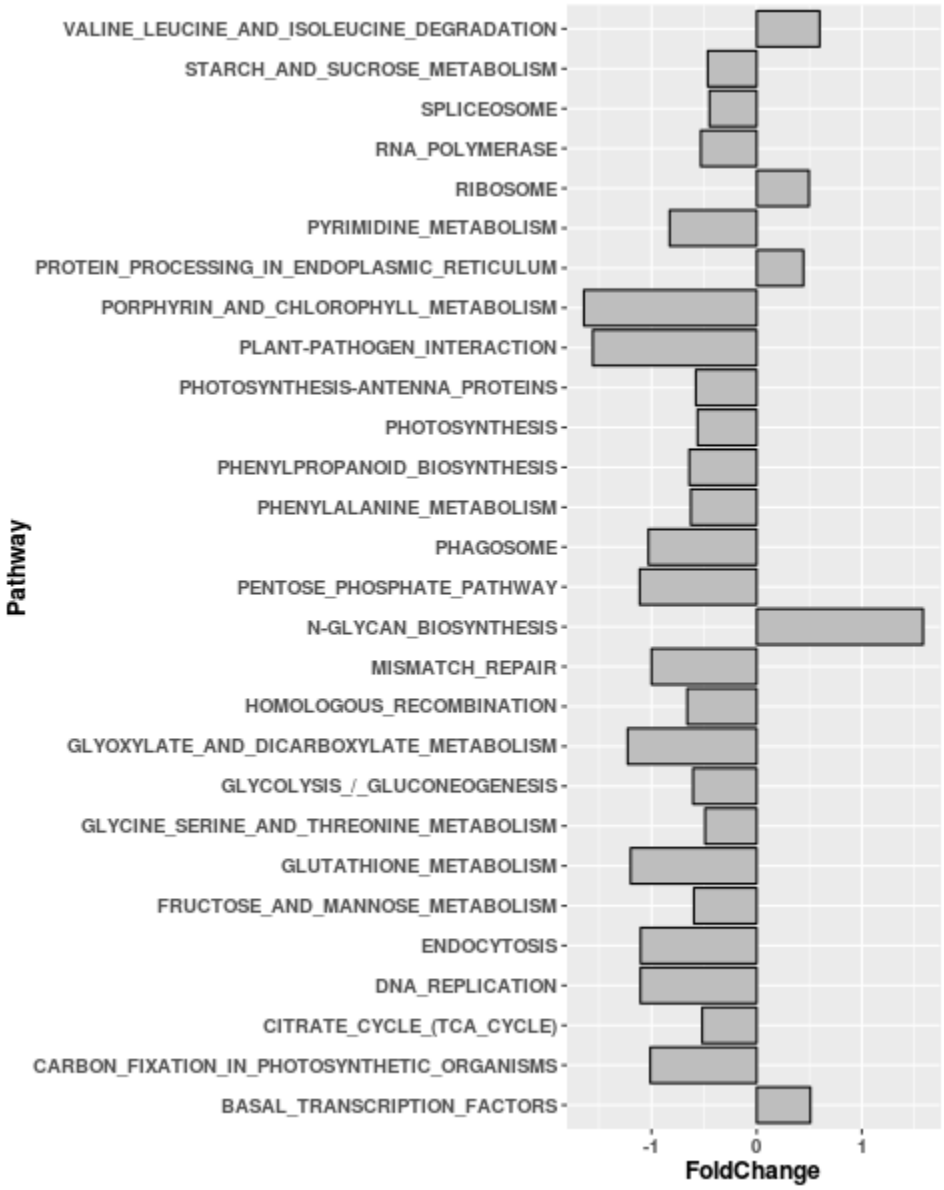

Figure 4.4: Bar plot depicting the fold change in density of KEGG pathways. The subnetwork density of each of the 117 KEGG pathways were computed in the control network as well as the drought network. Fold change was the calculated as $\log_2 (ND_d/ND_c)$, where ND_d and ND_c represents the density in the drought network and the control network, respectively, of each of the pathways labeled on the vertical axis.

Edge-set enrichment analysis using mean coexpression in the DDCN

The strength of coexpression among gene pairs within a given functional domain, was next evaluated for high or low significance, compared to all other domains in the DDCN. Specifically, the mean of coexpression within each functional class was computed, and divided by the standard deviation of the all mean scores to derive a Z score as the enrichment score. Following the central limit theorem, since the sampling size was large enough (about ~600 categories in GO BP terms and 117 KEGG pathways), the statistical significance could be estimated under the standard normal distribution, a procedure similar to the parametric analysis of gene-set enrichment algorithm (Kim and Volsky, 2005).

Under this framework, functional categories that had significantly higher or lower mean coexpression in the DDCN represented 30% of the 117 KEGG pathways (-log q-value > 2), and 12% of the 634 GO BP terms tested (-log q-value > 3). Among the KEGG annotations, the ribosome pathway was found to achieve the highest positive enrichment score (Table 4.2). The ribosome pathway is composed of genes engaged in the biogenesis of ribosomes which involves production of proteins and is responsible for growth. The second most cohesive pathway was the spliceosome, which is composed of 157 genes that largely participate in the removal of introns from mRNA and for generation of alternatively spliced isoforms, a phenomenon that is of much interest in plant abiotic-stress response research (Ding et al., 2014; Guerra et al., 2015; Thatcher et al., 2016). Of the 23 genes that function as splicing factors in the spliceosome pathway, 18 showed significantly altered expression levels in the drought microarray in at least one development stage of the three tested (Supplemental table 4.5). A closer examination by visualization of unfiltered coexpression between these splicing factors revealed that under drought, large alterations in the regulation of these genes occur, as most of the positively coexpressed

splicing factors in the CN switched to negative correlations in the DN (Fig. 4.5A). Similarly, the photosynthesis pathway also showed a large number of genes with altered correlation patterns (Fig. 4.5B). Interestingly, in the photosynthesis pathway, an iron-sulfur binding domain containing protein (LOC_Os01g64120) had an apparent increase in the number of negative correlations in the DN than the CN. This gene is a homolog of the recently discovered OsFdC2 gene that functions in the photosynthesis pathway by regulating electron transfer and chlorophyll content in rice (Zhao et al., 2015).

In addition to ribosome and spliceosome pathways, most KEGG pathways that are composed of genes encoding parts of a large protein complex also appear to be significantly enriched, and these correspond well with the GO BP terms that were found enriched in the edges of DDCN (Supplemental table S4.6). Altogether, the enrichment results point toward the fact that water stress has considerable effect on the 'translatome', as observed with other abiotic stresses (Sormani et al., 2011; Wang et al., 2013). Additionally, as expected, the cell-wall related GO BP terms were also found significantly enriched in the DDCN. The implication of cell wall metabolism under abiotic-stress is well known (Tenhaken, 2014), such as the role of pectins in drought tolerant wheat (Leucci et al., 2008).

Table 4.2: GO BP terms enriched in the DDCN. The mean coexpression score of each gene set in the GO BP annotations was divided by the standard deviation of the distribution to derive a Z score signifying whether the gene-set was significantly high or low in coexpression in the DDCN.

Pathway	# Genes	# Edges	Mean Score	Z Score
ribosome	217	7254	11.71	35.464
spliceosome	157	1255	13.421	19.555
RNA transport	118	768	13.567	15.62
oxidative phosphorylation	91	499	13.344	12.195
proteasome	58	341	13.651	10.531
ribosome biogenesis in eukaryotes	68	454	12.364	9.977
pyrimidine metabolism	87	255	13.246	8.594
purine metabolism	103	287	12.49	8.102
nucleotide excision repair	54	182	13.3	7.318
mRNA surveillance pathway	90	204	12.429	6.762
RNA degradation	75	165	12.949	6.61
protein processing in endoplasmic reticulum	148	240	11.826	6.593
aminoacyl-tRNA biosynthesis	54	195	12.342	6.514
phenylpropanoid biosynthesis	75	150	12.881	6.237
phagosome	62	108	12.965	5.361
ubiquitin mediated proteolysis	94	135	12.214	5.302
phenylalanine metabolism	65	96	12.881	4.99
mismatch repair	33	79	13.511	4.97
RNA polymerase	30	73	13.745	4.936
plant-pathogen interaction	73	66	13.401	4.472
DNA replication	43	186	10.536	4.409
base excision repair	33	59	13.591	4.344
homologous recombination	33	65	13.115	4.255
porphyrin and chlorophyll metabolism	33	71	12.62	4.117
endocytosis	78	90	11.872	4.072
basal transcription factors	33	47	13.14	3.632
citrate cycle (TCA cycle)	47	64	11.832	3.408
n-glycan biosynthesis	34	32	13.798	3.292
starch and sucrose metabolism	108	88	10.783	3.216
amino sugar and nucleotide sugar metabolism	84	44	12.503	3.179
glycolysis / gluconeogenesis	106	104	10.314	3.117
snare interactions in vesicular transport	33	19	14.85	2.9
photosynthesis-antenna proteins	14	65	1.915	-2.905
pyruvate metabolism	66	57	1.375	-3.044
photosynthesis	37	421	1.39	-8.248

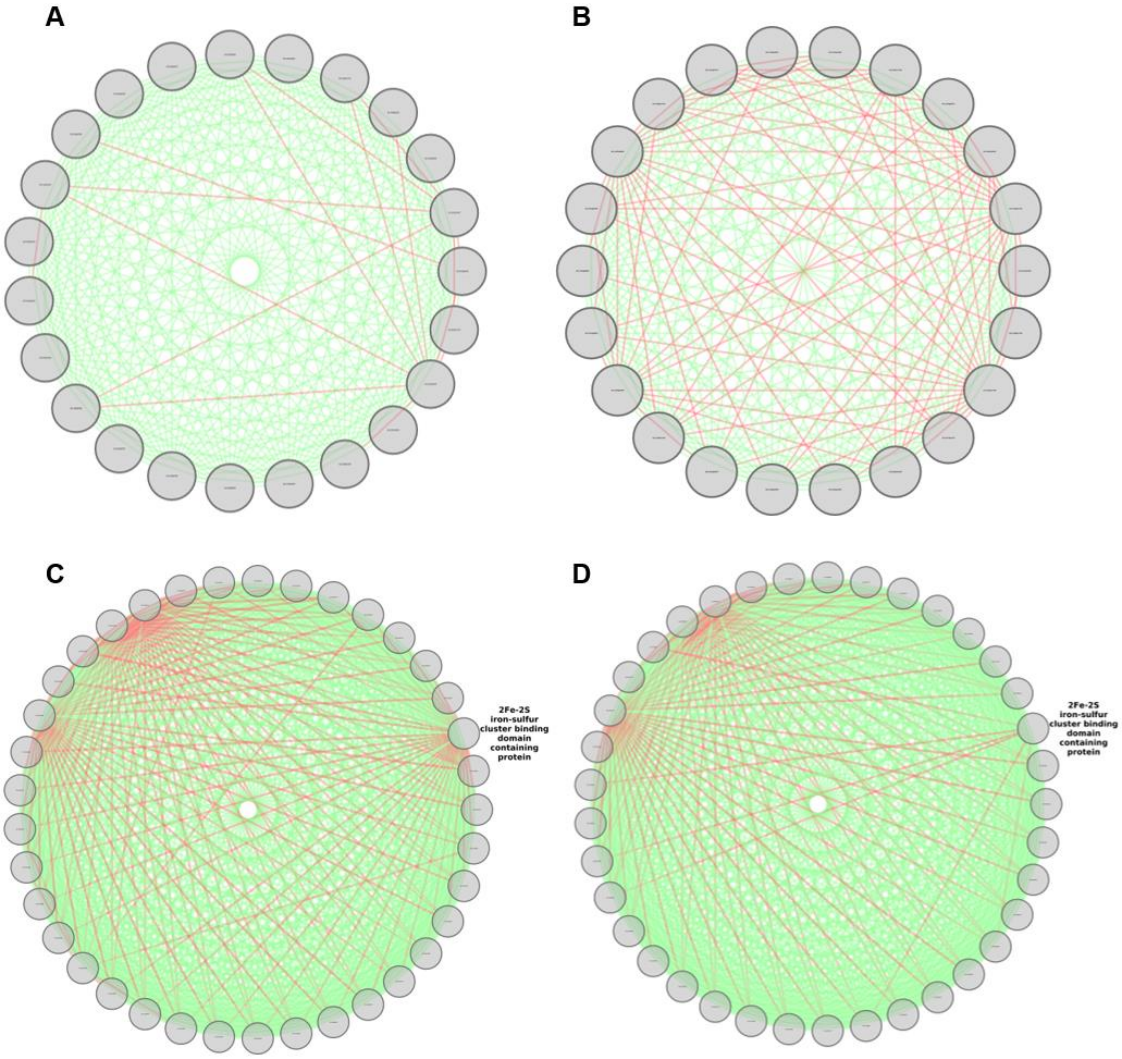

Figure 4.5: Rewiring of gene interactions in response to drought. Genes annotated as "splicing factors" were identified from the KEGG annotations and coexpression between them visualized using Cytoscape (Shannon et al., 2003) in the A) the control network and B) the drought network. Similarly, coexpression between the photosynthesis genes were visualized in C) the drought network and the D) the control network. Green and red lines (edges) connecting two genes represent positive and negative coexpression, respectively.

Discussion

The work described here used differential coexpression (DC) analysis to identify genes and pathways affected by water stress in rice. The method described deviates from the traditional norms of using a single expression dataset for identification of candidate genes based on differential expression (DE). Instead of focusing on the expression value of every gene individually, this method focuses on using the coexpression value of gene-gene pairs for analysis. Instead of DE, the gain or loss in correlation between gene pairs is calculated. The small overlap (maximum 30%) between 'drought responsive' genes identified through DE and DC analysis was surprising. The 4519 genes uniquely identified by DC represent genes that were not perturbed by drought, but still associated with genes that were affected. A stringent criteria to computationally evaluate whether these are truly drought associated or not would require evaluation of drought datasets from different genotypes, at least the ones that were represented in the underlying networks used for comparison. However, this evaluation was not possible at the current stage, as majorly all the available drought specific datasets were used in building the networks, and reusing the datasets would lead to correlated statistical biases and associated errors.

In an alternative approach to evaluate whether DC detected truly drought associated genes, the DCG set was evaluated computationally in a gene set enrichment analysis framework. The overlaps between gene lists from all the four sets (DCG and DE in seedling, vegetative and reproductive stages) and functional knowledgebase of rice (GO and KEGG annotations) were computed. The Fisher's exact test was used to evaluate the overlaps for statistical significance (Tian et al., 2017). This test revealed several drought associated pathways mentioned in the literature that could be recovered only from the enrichment analysis of DCG and not any of the other list of DE genes. However, a stringent experimental evaluation of the predictions would a

double blind test in which top *n%* genes are selected (based on *p*-values) and mixed with the same number of randomly selected genes from the pool of genes not detected by DC but only DE, as well as genes not detected by any method, and conducting a qPCR assay for drought response, possibly with different genotypes.

The limitation of this method of expression data analysis lies in the calculation of various enrichment tests. In parametric analysis (e.g. esea) it is assumed while calculating *p* values that the variables are independent of each other. This is obviously not true in the case of genes, as they depend on each other to function, and this becomes even more pronounced when dealing with edges. Whether this should be considered in a typical gene set enrichment analysis – given the computational costs involved in permutations and randomizations – has always been debated upon since the early days of GSEA and methods with similar concepts (Tamayo et al., 2016). However, it has also been shown that both parametric and non-parametric methods that consider the dependency between genes perform similarly and their outputs are not significantly different from each other (Kim and Volsky, 2005). In the light of this debate, the actual biological interpretation of gene set coherence in the model presented here is based on network densities, simply stating whether coexpression between member genes of a functional category increases or decreases under drought, as well as evaluation of coexpression strength between genes from different functional categories. This essentially means that in order to gain a strong biological inference from DC analysis, the underlying data sets used to create the coexpression networks should be as close to the biological condition in question as possible, even better if tissue specificity is accounted for.

The rewiring of biological processes and metabolic pathways in response to drought stress was examined on the basis of changes in the network densities. It was observed that drought triggers a massive system-wide transcriptional reprogramming, breaking communications between

genes functioning in same pathways while increasing communication between different pathways, indicating increased drought signaling. The loss of links in the DN was unexpected and counter-intuitive; a network under stress is expected to become more modular, meaning that communication between genes that participate in the same pathways is expected to increase and that between distantly related pathways is expected to decrease. This phenomenon has been observed in the coexpression network as well as the protein interaction network of yeast cells under oxidative stress (Lehtinen et al., 2013). However, in the analysis of rice datasets here, the opposite pattern was observed as the inter-pathway density increased and the intra-pathway density decreased in the DN as compared to the CN. This pattern could not be attributed to the selection of underlying datasets used for calculation of expression correlations, as several different control networks created using non-overlapping developmental phase datasets showed the same patterns of increase in inter-pathway and decrease in intra-pathway densities (data not shown). Overall, this indicated that in response to drought, the rice gene network significantly rewires and coordination of several pathways is required to overcome the stress, perhaps as a strategy of a coordinated stress resistance mechanism.

The change in network densities of biological processes and metabolic pathways has not been explored before in rice, as observed by our extensive literature survey. Hence, whether the observed patterns of pathway level cohesiveness can be generalized for all abiotic stresses or specific for drought stress remains untested, and calls for an in detail examination using datasets from different stresses in varying plant species. Nevertheless, the emergent properties of the rice drought network – that could not be determined using DE analysis – represent known mechanisms of drought response in rice and sheds light into several novel genes and pathways implicated under drought.

Methods

Estimation of coexpression

A total of six individual Affymetrix based rice expression datasets (GSE21651, GSE24048, GSE25176, GSE26280, GSE41647 and GSE81253) were used for the drought network (DN), and the dataset GSE19024 were used for the control network (CN). The data were downloaded from the Gene Expression Omnibus database (Edgar et al., 2002). Within each dataset, data were background corrected, normalized using the RMA algorithm in R (Irizarry et al., 2003) and the replicates averaged, yielding a total of 34 samples for the DN and 76 samples for CN. Individual probes were assigned to gene models using a custom CDF file of rice, retaining a total of 35,151 gene models for which expression could be estimated. To remove bias caused by genes expression at very low levels, both the expression matrices were filtered to retain only those genes that had an expression value above the 75th percentile of the dataset in at least one sample. Pearson's Correlation between the expression profiles of each gene-pair was calculated and Fisher's Z-transformed to a N(0,1) distribution (Soper et al., 1917). An absolute *Z* score of 2.58 was set as cutoff to declare a gene-pair as coexpressed (top 1% edges).

$$Z = \frac{1}{2}\, ln\, \frac{1+R}{1-R} \qquad (1)$$

Where R is the PC coefficient score of a gene pair.

Estimating Differential coexpression

Differential coexpression was computed as the difference in the absolute *Z*-scores of the static networks using the formula (de la Fuente, 2010; Jiang et al., 2016)

$$\Delta Z = \frac{|Z_d - Z_c|}{\sqrt{\frac{1}{N_d - 3} + \frac{1}{N_c - 3}}} \qquad (2)$$

Where Z_d and Z_c are the Z scores of the drought network and the control network, respectively, calculated using equation (1). N_d and N_c are the number of samples in the drought network and the control network, respectively. Since the ΔZ scores followed a Gaussian distribution, p-values were calculated under the standard normal distribution (Fukushima et al., 2012).

Estimating differential expression from microarrays

Raw data from GSE81253 were background corrected, normalized and summarized using the RMA algorithm in R (Gentleman et al., 2004). To reliably detect differential expression, genes with low variations were filtered if the IQR (interquartile range) was less than the median IQR. The limma model (linear models for microarray) was then used to detect differential expression of the retained genes (Smyth, 2004). The resulting *p*-values of the t-tests were converted to *q*-values to correct for multiple hypothesis testing (Storey and Tibshirani, 2003). Genes that had a *q*-value < 0.01 were declared as differentially expressed in each of the three samples.

Obtaining Molecular pathways and functional categories

Catalogs of genesets were downloaded from the PlantGSEA website. For GO ontologies, annotations from two servers were used. The first set was downloaded from the agbase server in September 2012 and was later updated in August 2016 with annotations from the plant GSEA server. Following transitive closure, genes in 'child' terms were annotated to all the 'parent'. Terms with less than 1500 annotations were retained to prevent statistical biases caused by non-

informative terms that annotate an overly large proportion of genes. Terms that annotated less than 10 genes were removed from statistical analysis. Since some of the terms in agbase are now considered obsolete, terms that were not present in the newer version of plantGSEA were discarded from the agbase server and the rest merged with plant GSEA server genesets. Further, terms that had very large overlaps in their corresponding annotations, estimated by calculating the Jaccard Index (JI) of overlap between all term-pairs, and removing the term with lower number of annotations for JI > 0.9 with a difference of at least 5 genes. This retained a total of 64312 annotations of 14384 genes by 683 specific and informative terms. For KEGG pathways, the top 3 terms of the canonical pathway categories were removed.

Parametric analysis of the edge-set enrichment of DDCN

The mean of coexpression within a given functional class (GO or KEGG) was calculated within each static network. A *Z* score was derived for each category as follows. First, from each static network (DN and CN), the mean of total coexpression scores (μ) and standard deviation of total coexpression scores (σ) was calculated. Then, for each functional class, the *Z* score was calculated as (Kim and Volsky, 2005)

$$Z = \frac{(Sm - \mu) * m^{\frac{1}{2}}}{\sigma} \qquad (3)$$

Where *Sm* is the mean of coexpression of the functional class (a gene set) and *m* is the size of the gene set. Both the networks were treated as fully connected where an edge not passing the significance threshold was taken as a *Z* of 0. *P*-values were obtained under the standard normal distribution and corrected for multiple testing with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

Calculation of Network Density

Density of a weighted undirected graph is calculated as $D = \frac{2 * \sum_{v \in V, u \in V, u \neq v} weight\ (u,v)}{V * (V-1)}$, where u and v are vertices (Liu et al., 2009), and weight in this context is the coexpression weight. Since this calculation required weights to range between 0 and 1, absolute values of Z scores of a given network (and subnetworks) were transformed to the required range. Calculations were done using a custom Perl script.

All network data was parsed using the Sleipnir library (Huttenhower et al., 2008), Network Analysis Tools (Brohee et al., 2008) and customized Perl and R scripts. Plots were made in R using the ggplot2 package (Wickham, 2009).

References

Atias O, Chor B, Chamovitz DA (2009) Large-scale analysis of Arabidopsis transcription reveals a basal co-regulation network. BMC Systems Biology **3:** 86-86

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. Nucl Acids Res **35**

Bassel GW, Lan H, Glaab E, Gibbs DJ, Gerjets T, Krasnogor N, Bonner AJ, Holdsworth MJ, Provart NJ (2011) Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. Proc Natl Acad Sci U S A **108:** 9709-9714

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol **57**

Bowne JB, Erwin TA, Juttner J, Schnurbusch T, Langridge P, Bacic A, Roessner U (2012) Drought Responses of Leaf Tissues from Wheat Cultivars of Differing Drought Tolerance at the Metabolite Level. Molecular Plant **5:** 418-429

Brohee S, Faust K, Lima-Mendez G, Sand O, Janky R, Vanderstocken G, Deville Y, van Helden J (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. Nucleic Acids Res **36:** W444-451

Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res **40:** D742-753

Cheadle C, Vawter MP, Freed WJ, Becker KG (2003) Analysis of Microarray Data Using Z Score Transformation. The Journal of molecular diagnostics : JMD **5:** 73-81

Chen G, Komatsuda T, Ma JF, Li C, Yamaji N, Nevo E (2011) A functional cutin matrix is required for plant protection against water loss. Plant Signaling & Behavior **6:** 1297-1299

Chia BKH, Karuturi RKM (2010) Differential co-expression framework to quantify goodness of biclusters and compare biclustering algorithms. Algorithms for Molecular Biology **5:** 23

Childs KL, Davidson RM, Buell CR (2011) Gene Coexpression Network Analysis as a Source of Functional Annotation for Rice Genes. PLoS ONE **6:** e22196

Choi Y, Kendziorski C (2009) Statistical methods for gene set co-expression analysis. Bioinformatics **25:** 2780-2786

Conway JR, Lex A, Gehlenborg N (2017) UpSetR: An R Package For The Visualization Of Intersecting Sets And Their Properties. bioRxiv

de la Fuente A (2010) From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet **26:** 326-333

Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L (2014) Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. BMC Genomics **15:** 431

D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics **16:** 707-726

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 36 (Database issue) **30**

Ferrero-Serrano Á, Assmann SM (2016) The α-subunit of the rice heterotrimeric G protein, RGA1, regulates drought tolerance during the vegetative phase in the dwarf rice mutant d1. Journal of Experimental Botany **67:** 3433-3443

Fracasso A, Trindade LM, Amaducci S (2016) Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE. BMC Plant Biology **16:** 115

Fukushima A (2013) DiffCorr: an R package to analyze and visualize differential correlations in biological networks. Gene **518:** 209-214

Fukushima A, Nishizawa T, Hayakumo M, Hikosaka S, Saito K, Goto E, Kusano M (2012) Exploring Tomato Gene Functions Based on Coexpression Modules Using Graph Clustering and Differential Coexpression Approaches. Plant Physiology **158:** 1487-1502

Gaiteri C, Ding Y, French B, Tseng GC, Sibille E (2014) Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav **13:** 13-24

Gao C, McDowell IC, Zhao S, Brown CD, Engelhardt BE (2016) Context Specific and Differential Gene Co-expression Networks via Bayesian Biclustering. PLOS Computational Biology **12:** e1004791

Gentleman RC, Carey VJ, Bates DJ, Bolstad BM, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth GK, Tierney L, Yang YH, Zhang J (2004) Bioconductor: Open software development for computational biology and bioinformatics. *In* Genome Biology,

Guerra D, Crosatti C, Khoshro HH, Mastrangelo AM, Mica E, Mazzucotelli E (2015) Post-transcriptional and post-translational regulations of drought and heat response in plants: a spider's web of mechanisms. Frontiers in Plant Science **6:** 57

Gupta C, Krishnan A, Collakova E, Wolinski P, Pereira A (2017) SANe: The Seed Active Network For Mining Transcriptional Regulatory Programs of Seed Development. bioRxiv

Han J, Shi X, Zhang Y, Xu Y, Jiang Y, Zhang C, Feng L, Yang H, Shang D, Sun Z, Su F, Li C, Li X (2015) ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. Scientific Reports **5:** 13044

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res **32**

Huttenhower C, Hibbs M, Myers C, Troyanskaya OG (2006) A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics **22:** 2890-2897

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

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249-264

Jiang Z, Dong X, Li Z-G, He F, Zhang Z (2016) Differential Coexpression Analysis Reveals Extensive Rewiring of Arabidopsis Gene Coexpression in Response to Pseudomonas syringae Infection. Scientific Reports **6:** 35064

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res **32**

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

Klebanov L, Jordan C, Yakovlev A (2006) A new type of stochastic dependence revealed in gene expression data. Stat Appl Genet Mol Biol **5:** Article7

Krishnan A, Gupta C, Ambavaram MMR, Pereira A (2017) RECoN: Rice Environment Coexpression Network for Systems Level Analysis of Abiotic-Stress Response. bioRxiv

Lai Y, Wu B, Chen L, Zhao H (2004) A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics **20:** 3146-3155

Lee B-R, Zaman R, Avice J-C, Ourry A, Kim T-H (2016) Sulfur Use Efficiency Is a Significant Determinant of Drought Stress Tolerance in Relation to Photosynthetic Activity in Brassica napus Cultivars. Frontiers in Plant Science **7:** 459

Lehtinen S, Marsellach FX, Codlin S, Schmidt A, Clement-Ziza M, Beyer A, Bahler J, Orengo C, Pancaldi V (2013) Stress induces remodelling of yeast interaction and co-expression networks. Molecular BioSystems **9:** 1697-1707

Leucci MR, Lenucci MS, Piro G, Dalessandro G (2008) Water stress and cell wall polysaccharides in the apical root zone of wheat cultivars varying in drought tolerance. J Plant Physiol **165:** 1168-1180

Lewis SE (2005) Gene Ontology: looking backwards and forwards. Genome Biol **6**

Li KC (2002) Genome-wide coexpression dynamics: theory and application. Proc Natl Acad Sci U S A **99:** 16875-16880

Li L, Briskine R, Schaefer R, Schnable PS, Myers CL, Flagel LE, Springer NM, Muehlbauer GJ (2016) Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. BMC Genomics **17:** 875

Liang Y-H, Cai B, Chen F, Wang G, Wang M, Zhong Y, Cheng Z-M (2014) Construction and validation of a gene co-expression network in grapevine (Vitis vinifera. L.). Horticulture Research **1:** 14040

Liany H, Rajapakse JC, Karuturi RKM (2017) MultiDCoX: Multi-factor Analysis of Differential Co-expression. bioRxiv

Liu G, Wong L, Chua HN (2009) Complex discovery from weighted PPI networks. Bioinformatics **25:** 1891-1897

Liu Y, Koyutürk M, Barnholtz-Sloan JS, Chance MR (2012) Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. BMC Systems Biology **6:** 65

Lui TWH, Tsui NBY, Chan LWC, Wong CSC, Siu PMF, Yung BYM (2015) DECODE: an integrated differential co-expression and differential expression analysis of gene expression data. BMC Bioinformatics **16:** 182

Ma C, Xin M, Feldmann KA, Wang X (2014) Machine Learning–Based Differential Network Analysis: A Study of Stress-Responsive Transcriptomes in Arabidopsis. The Plant Cell **26:** 520-537

Muthusamy M, Uma S, Backiyarani S, Saraswathi MS, Chandrasekar A (2016) Transcriptomic Changes of Drought-Tolerant and Sensitive Banana Cultivars Exposed to Drought Stress. Frontiers in Plant Science **7:** 1609

Peleg Z, Reguera M, Tumimbang E, Walia H, Blumwald E (2011) Cytokinin-mediated source/sink modifications improve drought tolerance and increase grain yield in rice under water-stress. Plant Biotechnol J **9:** 747-758

Pierson E, Koller D, Battle A, Mostafavi S, Ardlie KG, Getz G, Wright FA, Kellis M, Volpi S, Dermitzakis ET (2015) Sharing and Specificity of Co-expression Networks across 35 Human Tissues. PLoS Comput Biol **11:** e1004220

Rosa BA, Jasmer DP, Mitreva M (2014) Genome-Wide Tissue-Specific Gene Expression, Co-expression and Regulation of Co-expressed Genes in Adult Nematode Ascaris suum. PLOS Neglected Tropical Diseases **8:** e2678

Shaik R, Ramakrishna W (2013) Genes and Co-Expression Modules Common to Drought and Bacterial Stress Responses in <italic>Arabidopsis</italic> and Rice. PLoS ONE **8:** e77261

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**

Sircar S, Parekh N (2015) Functional characterization of Drought-responsive Modules and Genes in Oryza sativa: A Network-based Approach. Frontiers in Genetics **6**

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

Soper HE, Young AE, Cave BM, Lee A, Pearson K (1917) On the distribution of the correlation coefficient in small samples. appendixAppendix ii to the papers of "student" and r. a. Fisher. a cooperative study. Biometrika **11**

Sormani R, Masclaux-Daubresse C, Daniele-Vedele F, Chardon F (2011) Transcriptional Regulation of Ribosome Components Are Determined by Stress According to Cellular Compartments in Arabidopsis thaliana. PLOS ONE **6:** e28070

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

Stuart JM, Segal E, Koller D, Kim SK (2003) A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science **302**

Tamayo P, Steinhardt G, Liberzon A, Mesirov JP (2016) The Limitations of Simple Gene Set Enrichment Analysis Assuming Gene Independence. Statistical methods in medical research **25:** 472-487

Tenhaken R (2014) Cell wall remodeling under abiotic stress. Frontiers in Plant Science **5:** 771

Tesson BM, Breitling R, Jansen RC (2010) DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics **11:** 497

Thatcher SR, Danilevskaya ON, Meng X, Beatty M, Zastrow-Hayes G, Harris C, Van Allen B, Habben J, Li B (2016) Genome-Wide Analysis of Alternative Splicing during Development and Drought Stress in Maize. Plant Physiology **170:** 586-599

Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z (2017) agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Research **45:** W122-W129

Walley AJ, Jacobson P, Falchi M, Bottolo L, Andersson JC, Petretto E, Bonnefond A, Vaillant E, Lecoeur C, Vatin V, Jernas M, Balding D, Petteni M, Park YS, Aitman T, Richardson S, Sjostrom L, Carlsson LMS, Froguel P (2012) Differential co-expression analysis of obesity-associated networks in human subcutaneous adipose tissue. International journal of obesity (2005) **36:** 137-147

Wang J, Lan P, Gao H, Zheng L, Li W, Schmidt W (2013) Expression changes of ribosomal proteins in phosphate- and iron-deficient Arabidopsis roots predict stress-specific alterations in ribosome composition. BMC Genomics **14:** 783

Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer Publishing Company, Incorporated

Witt S, Galicia L, Lisec J, Cairns J, Tiessen A, Araus JL, Palacios-Rojas N, Fernie AR (2012) Metabolic and Phenotypic Responses of Greenhouse-Grown Maize Hybrids to Experimentally Controlled Drought Stress. Molecular Plant **5:** 401-417

Xu F, Yang J, Chen J, Wu Q, Gong W, Zhang J, Shao W, Mu J, Yang D, Yang Y, Li Z, Xie P (2015) Differential co-expression and regulation analyses reveal different mechanisms underlying major depressive disorder and subsyndromal symptomatic depression. BMC Bioinformatics **16:** 112

Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H (2014) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. **5:** 3231

Yuan L, Zheng C-H, Xia J-F, Huang D-S (2015) Module Based Differential Coexpression Analysis Method for Type 2 Diabetes. BioMed Research International **2015:** 836929

Zhang J, Li J, Deng H-W (2009) Identifying Gene Interaction Enrichment for Gene Expression Data. PLOS ONE **4:** e8064

Zhao J, Qiu Z, Ruan B, Kang S, He L, Zhang S, Dong G, Hu J, Zeng D, Zhang G, Gao Z, Ren D, Hu X, Chen G, Guo L, Qian Q, Zhu L (2015) Functional Inactivation of Putative Photosynthetic Electron Acceptor Ferredoxin C2 (FdC2) Induces Delayed Heading Date and Decreased Photosynthetic Rate in Rice. PLOS ONE **10:** e0143361

Chapter 5: Conclusions and General Discussions

The present study described the development of two bioinformatics applications for network-based predictions of plant gene functions. The Rice Regulatory Network (RRN) captures an abiotic-stress conditioned gene regulatory network designed to facilitate the identification of Transcription Factor (TF) targets that manifest during abiotic stresses. The network is essentially a consensus of predictions aggregated from three reverse-engineering algorithms applied to a large compendium of publicly available expression datasets in rice. The *Arabidopsis* Seed Active Network (SANe) is a gene regulatory network that encapsulates various aspects of seed formation, including embryogenesis, endosperm development and seed-coat formation. SANe is highly predictive of seed-specific gene function and further enhances our knowledge about regulatory mechanisms that underlie aforementioned processes of seed development. Both SANe and RRN are interactive web applications integrated with network analysis tools designed for use by experimental biologists.

Both the web tools introduce two new features to the current-state-of-art in gene network analysis methods: First, the enrichment analysis tool uses coexpressed gene clusters as gene sets to overcome the limitations of sparse functional annotations faced by knowledge-based approaches in interpreting genome-wide expression profiles. Second, apart from providing information about the biological processes and pathways perturbed in the uploaded transcriptome, the TFs that are statistically most relevant to the dataset are revealed. This result provides an intuitive framework for hypothesis generation and efficacious experimental design subsequently. In the development of these networks, some of the key observations are described as follows.

Integrating transcriptomes reveal much more than what was originally expected of each individual dataset. While a single microarray or RNA-seq experiment captures transcript accumulation from a static cellular state, an integrated analysis captures responses and variations

in a dynamic range of biological conditions, reflecting on the essentiality of genes, for example, in maintaining basic cellular processes (e.g. housekeeping genes). By design, microarrays cover about 65%-70% of all the identified gene models, limiting the genomic coverage in network inference. However, this can be overcome by integrating RNA-seq datasets that cover a larger fraction of genes and provide additional information like alternatively spliced isoforms. Statistical models for assembling a coexpression network using RNA-seq datasets are currently being developed and evaluated (Ballouz et al., 2015; Yu et al., 2017), and will be the standard norm for coexpression network inference in the future as the datasets accumulate in public repositories.

The choice of algorithms and statistical models for network inference available in published literature is large. As found by the DREAM challenge (Dialogue for Reverse Engineering Assessment and Methods) (Stolovitzky et al., 2007), an aggregate score of predictions from a large set of methods significantly increases the accuracy of network inference (Marbach et al., 2012). In chapter 2 – that is claimed to be the largest effort of integrating rice expression datasets till date – this study explored the concept of 'consensus network' on a compendium of 595 publicly available microarray samples of rice using Mutual Information (MI) as the underlying statistical model of coexpression. By evaluating and aggregating three algorithms and a Pearson's Correlation based method, it was shown that while the consensus network yields edges with larger overall accuracy scores, removing the worst performer increases this accuracy even more. In voting schemes, for example the Borda counts – a consensus approach similar to that described by DREAM – it is intuitive to think that larger the number of 'members' available for voting, larger will be the final accuracy of prediction. However, if the sampling of the members considered is biased, the accuracy will greatly vary. In the case of gene network inference, where a belief-based *gold standard* is usually available, the number of algorithms used in the ensemble is not the major

objective, but getting minimum false positives is. Hence, extensive evaluations of the consensus network is needed before proceeding with biological inferences from the network data. To make this process fast and easy, an attempt to automate the process will be made, and in future, the outcome will be released as a package in R (Team, 2012), with several reverse-engineering algorithms integrated into a single computational workflow.

Clustering is the most prominent step in network based assignment of gene functions. Several studies have used clustering methods that require a predefined number of clusters, for example in *k*-means clustering (Hartigan and Wong, 1979), the algorithm used for modeling uses a predefined parameter to obtain clusters, such as in the WGCNA methods (Langfelder and Horvath, 2008). However, it was observed that the parameter largely depends on the underlying datasets used for integration and the density of the obtained network. An extensive data-driven approach should be undertaken to define a clustering parameter that best reveals the clusters inherent in the network data. The method presented here used two popular clustering algorithms (SPICi and MCL) and evaluated the parameter required by these algorithms against a GO based *gold standard*. This led to the evaluation of a range of clustering parameters to determine the threshold which best preserves the functional and topological properties of the network.

At the present time, high dimensional data analysis in bioinformatics is a routine in molecular biology labs, and most often biologists are faced with little or no programming experience required to analyze such datasets. This calls for the development of more sophisticated computational workflows that can be easily integrated in labs with researchers who routinely analyze high throughput genomic datasets. To foster such an environment, the data and workflows developed during the entire course of the research presented here are made available in easy to use web-based platforms with manuals and demonstration links. Although the databases presented

here deal with two specific biological questions, seed development and stress response in plants, the range of biological queries that can be answered using the approaches devised here are widely applicable to other datasets with varying biological themes (e.g. biotic stress response, post-germination stage network) as well as on datasets from different organisms. In future, the methods implemented for the construction of the seed network described in chapter 3 will be re-used for different tissues/cell-types/organs of the Arabidopsis plant (e.g. roots, flowers and leaves) and presented to the user with a choice of networks to select and analyze their dataset.

Functional interactions between genes is the fundamental principle of metabolic pathways, and this phenomenon is overlooked in current state of gene expression analysis methods that consider expression variation of every gene individually. For example, non-heritable expression changes arising due to epigenetic factors (Seo et al., 2016), genes that do not perturb in expression but have a significant role in the underlying phenotypes, or genes that do change in expression but the magnitude is so small that they fail to pass the user-defined statistical thresholds (Cheng et al., 2004). In a 'differential networking' analysis framework, a novel network-density based algorithm is proposed to detect gain or loss in correlation amongst genes annotated to specific functional domains within two conditionally independent networks. The algorithm leverages on known pathway genes and two edge-weighted gene networks that it takes as input. The output is a list of genes with significantly rewired connections and a list of pathways that have significantly changed their cohesiveness between the two input networks. The workflow was tested on control and drought coexpression networks of rice, showing that several well-known drought response pathways could not be detected on the basis of differential expression, but resurfaced only in the light of differential coexpression.

Integration of coexpression networks with GWAS data

The current state of outcome of a Genome Wide Association Study (GWAS) is a set of SNPs that reach a genome-wide significance level, which requires a stringent p-value cutoff. This strategy of deriving associations excludes many genuinely associated SNPs that have a weak or moderate association signal with the trait of interest, hence overlooking joint effect of multiple SNPs/genes. In humans, gene set enrichment analysis has been used to boost the power of GWAS data (Wang et al., 2010; Weng et al., 2011). These gene sets come in the form of ontologies such as the Gene Ontology and the Trait Ontology and canonical pathways such as the CYC pathways, KEGG pathways and Mapman terms (Ashburner et al., 2000; Ware et al., 2002; Kanehisa et al., 2004; Thimm et al., 2004). In plants, the information in these pathways has been increasing rapidly with frequently sequenced newer genomes, and completely new ontologies are being developed (e.g. abiotic stress ontology). Several statistical models and algorithms have been brought forth that can be used to overlay this pathway-level information onto the GWAS data, essentially revealing those variants/genes that are functionally coupled and jointly associated with the trait of interest, hence keeping the false negatives to a minimum.

One of the common limitation of gene set enrichment analysis is that the current knowledgebase regarding functional annotations of genes has many missing parts. This sparsity in functional annotations is overcome by network-guided approaches to increase the flexibility of genesets. Network data, such as the protein-protein interaction network, coexpression networks and regulatory networks can be used as a reference network, and p-values of gene-SNP associations can be directly superimposed on the network. The jActiveModules plugin (Trey Idekar Lab, UCSD) available in Cytoscape (Shannon et al., 2003), directly uses the network topology and association p-values of each gene to extract meaningful ‘active modules’ of highly

interconnected genes that are also associated with relevant SNPs for a trait of interest. This algorithm has been applied to a human multiple sclerosis GWAS data using a PPI network as guide (Baranzini et al., 2009). In plants, a few studies have examined the integration of genomic data with gene coexpression data to find small effect genes for traits, such as for Glucosinolates in Arabidopsis (Chan et al., 2011), seed development traits in *Brassica napus* (Korber et al., 2015), and using publicly available SNPs for traits in rice (Ficklin and Feltus, 2013).

References

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. Nat Genet **25**

Ballouz S, Verleyen W, Gillis J (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics **31:** 2123-2130

Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BMJ, Kappos L, Gene MSAC, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Human Molecular Genetics **18:** 2078-2090

Chan EK, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in Arabidopsis thaliana. PLoS Biol **9:** e1001125

Cheng C, Pounds SB, Boyett JM, Pei D, Kuo ML, Roussel MF (2004) Statistical significance threshold criteria for analysis of microarray gene expression data. Stat Appl Genet Mol Biol **3:** Article36

Ficklin SP, Feltus FA (2013) A systems-genetics approach and data mining tool to assist in the discovery of genes underlying complex traits in Oryza sativa. PLoS One **8:** e68551

Hartigan JA, Wong MA (1979) Algorithm AS 136: a K-means clustering algorithm. J Royal Stat Soc Series C (Applied Statistics) **28**

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res **32**

Korber N, Bus A, Li J, Higgins J, Bancroft I, Higgins EE, Parkin IA, Salazar-Colqui B, Snowdon RJ, Stich B (2015) Seedling development traits in Brassica napus examined by gene expression analysis and association mapping. BMC Plant Biol **15:** 136

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9:** 559

Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. Nat Meth **9:** 796-804

Seo M, Kim K, Yoon J, Jeong JY, Lee H-J, Cho S, Kim H (2016) RNA-seq analysis for detecting quantitative trait-associated genes. Scientific Reports **6:** 24375

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**

Stolovitzky G, Monroe D, Califano A (2007) Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. Ann N Y Acad Sci **1115:** 1-22

Team RDC (2012) R: A Language and Environment for Statistical Computing. the R Foundation for Statistical Computing.

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. Plant J **37:** 914-939

Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet **11:** 843-854

Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S, McCouch S, Stein L (2002) Gramene: a resource for comparative grass genomics. Nucleic Acids Res **30**

Weng L, Macciardi F, Subramanian A, Guffanti G, Potkin SG, Yu Z, Xie X (2011) SNP-based pathway enrichment analysis for genome-wide association studies. BMC Bioinformatics **12:** 99

Yu H, Jiao B, Liang C (2017) High-Quality Rice RNA-Seq-Based Co-Expression Network For Predicting Gene Function And Regulation. bioRxiv