USING TRANSCRIPTOME AND DATA SCIENCE METHODS TO UNCOVER GENE REGULATORY AND FUNCTIONAL INFORMATION

By

Sahra Uygun

A DISSERTATION

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

Genetics – Doctor of Philosophy

2017

ABSTRACT

USING TRANSCRIPTOME AND DATA SCIENCE METHODS TO UNCOVER GENE REGULATORY AND FUNCTIONAL INFORMATION

By

Sahra Uygun

Even in the well-studied model organisms, there are still genomic regions with unknown function. These genomic regions include protein-coding genes and regulatory elements that are key components of transcriptional regulation. With technological advances, more biological data are being generated including spatial, temporal, developmental, and conditional gene expression data. Gene expression data, and specifically co-expression analyses have been widely used to predict gene function through guilt by association. However, it remains to be seen to what degree co-expression is informative, whether it can be applied to genes involved in different biological processes, and how the choice of gene expression dataset and clustering algorithms impact inferences about gene functions. To answer these questions, I used co-expression to identify novel genes that function in a biological process, and the impact of different clustering algorithms on the ability to identify genes that function in the same pathway. Apart from the functional associations, gene co-expression analyses can also be used to identify the putative *cis*-regulatory elements that are over-represented in co-expressed gene promoters. These elements can be used to build models of gene regulation under changing environments and genome-wide models of how different organ and cell type gene expression are regulated under changing environments have not yet been built in plants. I used *Arabidopsis thaliana* organ and cell type stress responsive gene expression data and co-expression clusters to identify putative *cis*-regulatory elements. Using these elements and machine learning models, I predicted high salinity responsive gene expression in shoots, roots and six root cell types. I found that plant organ and cell type transcriptional response to high salinity are likely regulated by a core set of elements that we identified and built predictive models of plant spatial transcriptional responses to environmental stress. Overall, this research contributes to understanding the role of “big data” in biology, provides guidelines for effectively using gene co-expression in functional associations and shows how computational approaches help in identifying gene regulatory information.

ACKNOWLEDGEMENTS

I would like to thank my advisor Shin-Han Shiu for his endless support and patience throughout my PhD. With his mentorship, I learned how to ask constructive scientific questions and approach them. He always created a fun and active environment in his laboratory. I would also like to thank my committee members, who all helped me in my PhD with their suggestions and feedback. I was fortunate to rotate in Robert L. Last's laboratory during my first year of PhD. Later, he became one of my committee members and collaborator mentor in various projects and helped me with his scientific expertise and critical thinking. I was also so lucky to rotate in Jin Chen's laboratory during the first year. With his computational biology expertise, I was inspired to use computational approaches to answer biological questions and he was always ready to help. In 2011, I took Christina Chan's Systems Biology course, where I was introduced to various topics in functional genomics. I was so happy when she agreed to be on my committee. I am so grateful for her suggestions and support. Apart from my committee members, other professors also played a huge role in my PhD. I cannot end my acknowledgements without mentioning Cathy Ernst, Genetics Graduate Program director. She was always there for the Genetics students and was so helpful. It was so good to know that we had her support in all aspects of PhD. I would also like to thank the members of the Shiu laboratory over the years. The laboratory grew in numbers and members changed while I was working there and each member was valuable for both scientific discussions and friendship. I would like to specifically thank Melissa Lehti-Shiu for her endless support. Finally, without the continuous encouragement and love of my family, I would not be to complete my PhD. I would like to thank Ayşegül and Öner Uygun for always being there for me and I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF FIGURES

KEY TO ABBREVIATIONS

ABA	Abscisic Acid
AlignACE	Aligns Nucleic Acid Conserved Elements
AUC-ROC	Area Under Curve - Receiver Operating Characteristic
BIC	Bayesian Information Criterion
BN	Bayesian Network
cDNA	Complementary DNA
CFS	Correlation-based Feature Selection
ChIP	Chromatin Immunoprecipitation
CNS	Conserved Non-coding Sequence
COL	Columella
COR	Cortex
CRC	cis-regulatory Code
CRE	cis-regulatory Element
DAP	DNA Affinity Purification
DHS	DNase I Hypersensitive Sites
DNA	Deoxyribonucleic Acid
EBI	European Bioinformatics Institute
EC	Expression Coherence
END	Endodermis
EPI	Epidermis
eQTL	Expression Quantitative Trait Loci

FET	Fishers Exact Test
FRiP	Fraction of Reads in Peaks
GEO	Gene Expression Omnibus
GFP	Green Fluorescent Protein
GO	Gene Ontology
GO-BP	Gene Ontology Biological Process
IUPAC	The International Union of Pure and Applied Chemistry
KL	Kullback-Leibler
LeuDeg	Leucine Degradation Pathway
LOR	Log-odds Ratio
MDScan	Motif Discovery Scan
MEME	Multiple Expectation Maximization for Motif Elicitation
MI	Mutual Information
NASCArray	Nottingham Arabidopsis Stock Centre Array
NCBI	The National Center for Biotechnology Information
PBM	Protein Binding Array
PCC	Pearson Correlation Coefficient
pCRE	Putative CRE
PHL	Proto-phloem
PWM	Position Weight Matrix
RBF	Radial Basis Function
RF	Random Forest
RMA	Robust Multi-Array Average

RNA	Ribonucleic Acid
SAGE	Serial Analysis of Gene Expression
STE	Stele
SVM	Support Vector Machine
TAIR	The Arabidopsis Information Resource
TAMO	Tools for Analysis of Motifs
T-DNA	Transfer DNA
TF	Transcription Factor
TFBM	Transcription Factor Binding Motif
WEKA	Waikato Environment for Knowledge Analysis
WGCNA	Weighted Gene Co-expression Network Analyses
WT	Wild-type
YMF	Yeast Motif Finder

CHAPTER 1

INTRODUCTION

1.1 Omics data and its use in functional genomics

The term “omics” is commonly used in biological sciences to refer to genome-scale data [1]. For example, genomics, first used by Thomas Roderick in 1986, refers to sequencing and analyzing the genomes of organisms [2]. Currently 3,808 eukaryotic genomes, including more than 100 plant species [3], are available in The National Center for Biotechnology Information (NCBI) database [4]. Even though genome sequences are useful, they only constitute a “natural coordinate system” [5], where the individual components involved in how organisms develop, function and respond to the environment are left to be identified. Once a genome has been sequenced, the next step is to identify and annotate the functional regions in the genome. This has been the major focus of functional genomics approaches, which have the goal of understanding functions of unknown genes and other regions in genome to understand how organisms function [5,6]. Annotations are descriptions of genomic features, and these descriptions can be structural or functional [7]. Structural annotations include specifying the coordinates, splice forms, intron/exon junctions of genes and could also include regulatory site information. Structural annotation of genes can be considered as the starting point for characterization of gene functions [3]. After structural annotation of genes is carried out, cellular function and location of gene products as well as the biological processes that the genes are involved in can be characterized. All these biological roles can be considered as functional annotations [8]. One of the common approaches towards annotating biological roles of genes is using comparative sequence analyses, comparing sequence of unknown genes to known ones [9], particularly through interspecies comparisons. Sequence-based functional inference relies on high sequence similarity, which might reflect functional similarity but not all functionally related genes share sequence similarity [10].

Genomic sequence alone is not sufficient to pinpoint biologically meaningful regions. To uncover gene function, high-throughput data have been generated covering multiple layers of biological information involving chromatin, DNA, RNA, proteins and metabolites [11]. For example, chromatin level information includes nucleosome occupancy and histone modifications, while genome level information includes genomic sequences and genome binding by regulatory components like transcription factors (TFs). Gene product information such as quantification of transcripts, proteins, and metabolites have been generated at a genome-scale. Different levels of biological information contribute to understanding functions of unknown genes that are the goals of functional genomics approaches. At the levels of gene products, transcript level information —i.e. steady-state mRNA levels— is the most abundant data type [12] and available from publicly data repositories such as NCBI. This facilitates the computational studies to utilize the vast number of gene expression data to extract biological information.

1.2 Abundance and utility of gene expression profiling data

Gene expression data are useful either alone or in combination with other data for disease classification [13–16], marker gene discovery [16], expression quantitative trait loci (eQTLs) discovery [17–19], learning about gene evolution [20], and identifying gene co-expression networks that can lead to gene regulatory and functional information [9,21,22]. In achieving these, publicly available gene expression datasets are useful resources. Gene Expression Omnibus (GEO, [23]) is a public functional genomics data repository [24]. Originally, this repository is started for storing microarray and sequencing based high-throughput gene expression data similar to the ArrayExpress repository of European Bioinformatics Institute (EBI, [25]). However with technological advances, other biological data, such as nucleosome occupancy, are also being stored there. Nonetheless, gene expression profiling is the most abundant data type in this repository. Gene expression profiles could be obtained by microarray [26] and RNA-seq [27] experiments. These methods are the most widely used gene expression measurement approaches. For example, 48,501 gene expression datasets available in GEO are based on microarray technology. Microarrays for detecting gene expression was developed in 1995 [28] and the aim was to monitor the expression of many genes quantitatively in parallel [26]. This technique involves fixed DNA probes that the cDNA from biological sample can bind to and hybridize [26]. As hybridization is involved, using microarrays require prior knowledge of genome sequence, which is a limitation if an organism of unknown genome or transcriptome is studied. Also, cross-hybridization and saturation of hybridization are additional limitations of array-based approaches in monitoring gene expression [29]. To overcome some of these limitations, sequencing-based techniques have been developed, where prior sequence information is not required [27]. Currently, 11,338 gene expression datasets available in GEO are based on high-throughput sequencing. Different from array-based approaches, cDNA sequence can be directly determined using sequencers. Sequencing has also been improved from traditional Sanger sequencing and tag-based sequencing (such as Serial Analysis of Gene Expression (SAGE) [30]) to next-generation sequencing [29].

Through the genome-scale expression profiling techniques, it is possible to understand which genes are induced/repressed and highly/lowly expressed during a developmental stage, in a particular tissue, and under a particular condition. However, the number of genes altered in gene expression is not the only information that could be obtained from systems-wide gene expression data, as it is possible to monitor thousands of genes and compare multiple gene expression profiles. This has led to substantial efforts in the field of bioinformatics to advance and develop data mining strategies to use the accumulating gene expression data to identify functions of genomic regions. For example, co-expression analysis relies on the observation that functionally related genes often have similar expression patterns and can be useful in reverse genetics approaches for narrowing down candidate genes to test for a particular function [10]. In addition, co-expressed genes tend to share common regulatory signatures such as similar TF binding sites [31]. This also makes gene expression data useful in finding over-represented motifs (potential TF binding sites) from co-expressed genes. Databases of co-expression gene networks, where expression similarity between genes (nodes) are captured in connections (edges), are available for multiple plant species (e.g. *Arabidopsis thaliana*, rice, barley and others [32–35]) as well as other organisms like fruit fly, mouse, and humans [36].

In the following sections, I discuss the utility of using gene expression data (temporal, spatial, and conditional) and co-expression analyses in hypothesizing gene function as well as identifying potential TF binding sites. It should be noted that the gene expression data only reflects one level of gene regulatory information among multiple levels of biological information mentioned earlier. Through integration of different levels of information, it is possible to get a more complete picture of inferred "function" compared to using one data type. This type of data integration efforts have been carried out in multiple organisms including *Arabidopsis thaliana*, fruit fly, mouse, and humans [37–39] by combining transcriptome with, for example, protein-DNA interactions, protein-protein interactions, and literature co-occurrences [40]. However, in data integration approaches, transcriptome data remain the most abundant and the most influential in capturing gene functional relationships [10,41].

1.3 Assigning functions to genes via computational approaches using gene expression data

Gene function can be characterized in multiple ways. Function of a protein-coding gene can be described by the biological processes and pathways its products are involved in, the molecular functions its gene product have, and/or the cellular component the gene product is located in. These properties of genes are summarized in controlled vocabulary in Gene Ontology (GO), where ~20 organisms have gene annotations to one of the following hierarchical ontologies: biological processes, molecular function and cellular component [42]. In the model plant, *Arabidopsis thaliana*, 40% of protein-coding genes have annotations with experimental evidences in at least one of the ontologies [10]. However, only ~5% of *A. thaliana* genes have annotations to all ontologies based on experimental evidence [10]. In addition, it is remarkable that only 1% of rice protein-coding genes have annotation in at least one of the ontologies based on experimental evidence [10]. This leaves over 90% of genes to characterize experimentally. Thus, it is necessary to computationally predict gene function and narrow down the candidate genes that would be further experimentally tested.

Gene expression data are useful in hypothesizing and predicting gene function. If a gene X with unknown function coordinately expresses over a variety of experimental conditions (co-expresses) with a gene Y of known function, then it is possible that the gene X has a similar function as Y. This is known as guilt by association [43]. The relationship between co-expression and function was first demonstrated in *Saccharomyces cerevisiae* and human transcriptome studies [45–48]. This approach has been used in plants as well to identify genes that are involved in multiple pathways including fatty acid biosynthesis, specialized metabolism, and secondary cell wall biosynthesis [10,49–51]. For example, 71 co-expression modules that were associated with genes that are expressed in specific tissues or in response to pathogen infection, abiotic stress, hormone treatments or environmental conditions were identified in *Oryza sativa* (rice) [52]. Importantly, 17% of the 17,298 rice genes that lacked a functional description (GO annotation) were found in at least one of the co-expression modules that could be functionally associated to other genes in the module [52]. In another recent study, researchers used co-expression to hypothesize gene function using guilt by association in *Bos taurus* (cattle) and they further supported their predictions with protein interaction data. Overall, 132 genes with previously unknown function were assigned biological roles [53].

Although co-expression is useful in predicting gene function, there are limitations to consider. First, assessing co-expression is not trivial. How similarity of expression is calculated and how high similarity is defined are expected to impact which genes are considered to be co-expressed. Thus, their definition can change conclusions regarding gene functional associations. The second limitation is that, depending on the number and type of gene expression samples used in co-expression analysis, the identity of genes defined as co-expressed might change. Current studies typically combine multiple datasets for gene function inference [9,25]. Including large numbers of samples (e.g. over ~100s) increases the statistical power of calculating expression similarity. However, combining gene expression samples from different experiments may result in the loss of context-specific relationships [54]. For example, similarities in gene expression profiles may depend on the cellular context [55], such as the differences in co-expression groups in cancer vs. non-cancer cells [56]. The inclusion of additional data may eliminate the co-expression signal. The third limitation of using co-expression in functional inference is that the genes of interest might be regulated at a level other than transcription, such as post-transcriptional regulation. In this case, the gene co-expression relationship might not be informative to pinpoint functionally related genes and data integration methods that incorporate epigenetic and post-transcriptional information are expected to be more informative compared to using co-expression alone. Apart from gene functions, *cis*-regulatory code can be inferred from analyzing co-expressed genes. The knowledge of components of gene regulatory machinery is important for understanding how genes are regulated at the transcriptional level. Among the regulatory components, the *cis*-regulatory elements (CREs) that the TFs bind to drive gene expression are important for regulating temporal, spatial and conditional gene expression. Gene expression data and co-expression analyses are useful in identifying CREs. CREs can then be used to form predictive models of gene expression.

1.4 Deciphering key players in gene expression regulation

Gene regulation ensures that the gene products are made correctly in a temporal, spatial and conditional manner [57,58] and involve multiple levels. The multiple levels of regulation leading to the final protein products include epigenetic, transcriptional, post-transcriptional, and post-translational level. Steady-state mRNA levels are particularly determined by transcriptional regulation [59]. At the transcriptional level, CREs and the DNA-binding TFs are important components that recruit the basal transcriptional machinery including RNA polymerase.

One way to identify CREs is to use chromatin immunoprecipitation (ChIP) with a given TF followed by array hybridization or sequencing, which yields binding site sequences [60]. In addition to identifying TF binding sites with ChIP studies, chromatin features such as histone modifications can also yield gene regulatory information. For example, ChIP-seq focusing on specific histone modifications (e.g. H3K27 acetylation and H3K4 methylation [61]) and open chromatin regions identified from DNAse I hypersensitivity [62] can provide information on potential CREs. Another way to identify potential CREs is using computational methods to identify over-represented sequences among a given set of sequences. CREs are often found close to transcription start site of the genes [63,64] and using genomic sequences, particularly promoter regions of the genes that co-express over a variety of conditions [65,66] it is possible to identify potential TF binding elements. Other genomic regions, such as the first intron of the gene itself, are also found to contain CREs involved in gene regulation [67].

It should be noted that the above approaches offer genome-wide information on TF binding and/or CREs. However, they do not provide mechanistic details of transcriptional regulation of genes. To further complement the genome-wide approaches mentioned above, machine learning techniques can be used to predict gene expression based on rules of CREs. The rules can be based on presence/absence, copy number, location, combinations of CREs. Machine learning includes statistical modelling [68] and has three stages: algorithm design, learning and testing. Potential CREs identified from genome-wide approaches can be used to build models for predicting gene expression and, based on the performance of prediction models, a set of CREs might be determined as drivers of gene expression at the context studied. Thus, these data-driven methods are key tools for the identifying CREs by forming predictive models [70].

1.5 Thesis chapters

In this research, I used publicly available high-throughput gene expression data to identify functional associations and putative CREs involved in organ and cell type stress responsive gene expression. In research described in Chapter 2, I investigated the utilities and limitations of using gene co-expression in hypothesizing functional association. Specifically, I assessed to what extent *Arabidopsis thaliana* pathway genes co-express. I also determined what type of pathways tend to form co-expression modules and evaluated the influence of dataset on the co-expressed genes. I also evaluated the impact of commonly-used clustering algorithms and their parameters on the ability to identify genes that function in the same pathways. In validating co-expression cluster memberships, I used an independent phenomics dataset to confirm the potential functional associations obtained from clusters. In Chapter 3, I explored what the *cis*-regulatory code is for the organ specific high-salinity responsive gene expression. I asked to what extent the stress gene expression is similar among different organs and whether the current knowledge of TF binding sites could explain the organ-specific stress gene expression. Through collaboration with Alexander E. Seddon, who completed his Master's study in the Shiu laboratory, we identified putative CREs that might be involved in organ high-salinity responsive gene expression and formed predictive models of gene up-regulation. I incorporated known TF binding sites, chromatin accessibility and evolutionary conservation; Alexander E. Seddon incorporated CRE combinatorial relations in machine learning models. Overall, we present a genome-wide view of *cis*-regulatory logic of organ gene expression in response to high-salinity. In research described in Chapter 4, I asked whether whole organ associated CREs could explain gene expression at a finer resolution to cell type level. I identified root cell type putative CREs and formed predicted models for each cell type.

REFERENCES

REFERENCES

1. Joyce AR, Palsson BØ. The model organism as a system: integrating “omics” data sets. Nat Rev Mol Cell Biol. Nature Publishing Group; 2006;7: 198–210. doi:10.1038/nrm1857

2. Hieter P, Boguski M. Functional Genomics: It’s All How You Read It. Science (80-). 1997;278.

3. Veeckman E, Ruttink T, Vandepoele K. Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. Plant Cell. 2016;28: 1759–68. doi:10.1105/tpc.16.00349

4. National Center for Biotechnology Information. [cited 14 Dec 2016]. Available: https://www.ncbi.nlm.nih.gov/

5. Werner T. Next generation sequencing in functional genomics. Brief Bioinform. Oxford University Press; 2010;11: 499–511. doi:10.1093/bib/bbq018

6. Holtorf H, Guitton M-C, Reski R. Plant functional genomics. Naturwissenschaften. 2002;89: 235–49. Available: http://www.ncbi.nlm.nih.gov/pubmed/12146788

7. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18: 188–196. doi:10.1101/gr.6743907

8. Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, et al. Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol. American Society of Plant Biologists; 2004;135: 745–55. doi:10.1104/pp.104.040071

9. Movahedi S, Van bel M, Heyndrickx K s., Vandepoele K. Comparative co-expression analysis in plant biology. Plant Cell Environ. Blackwell Publishing Ltd; 2012;35: 1787–1798. doi:10.1111/j.1365-3040.2012.02517.x

10. Rhee SY, Mutwil M. Towards revealing the functions of all genes in plants. Trends Plant Sci. Elsevier Ltd; 2014;19: 212–221. doi:10.1016/j.tplants.2013.10.006

11. Berger B, Peng J, Singh M. Computational solutions for omics data. Nat Rev Genet. NIH Public Access; 2013;14: 333–46. doi:10.1038/nrg3433

12. Barrett T. Gene Expression Omnibus (GEO). The NCBI Handbook. National Center for Biotechnology Information (US); 2013.

13. van ’t Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415: 530–536. doi:10.1038/415530a

14. Weigelt B, Baehner FL, Reis-Filho JS. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. J Pathol. 2010;220: 263–80. doi:10.1002/path.2648

15. Yeoh E, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell. 2002;1: 133–43. doi:10.1016/S1535-6108(02)00032-6

16. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science (80-). 1999;286.

17. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. 2015;16: 197–212. doi:10.1038/nrg3891

18. Ranjan A, Budke JM, Rowland SD, Chitwood DH, Kumar R, Carriedo L, et al. eQTL Regulating Transcript Levels Associated with Diverse Biological Processes in Tomato. Plant Physiol. 2016;172: 328–40. doi:10.1104/pp.16.00289

19. West MAL, Kim K, Kliebenstein DJ, Van Leeuwen H, Michelmore RW, Doerge RW, et al. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics. 2007;175: 1441–1450. doi:10.1534/genetics.106.064972

20. Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC Evol Biol. 2011;11: 47. doi:10.1186/1471-2148-11-47

21. Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF, et al. Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol. 2008;147: 41–57. doi:10.1104/pp.108.117366

22. Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y. Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol. 2009;150: 535–46. doi:10.1104/pp.109.136028

23. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. Oxford University Press; 2002;30: 207–10. doi:10.1093/NAR/30.1.207

24. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;14: 89–99. doi:10.1038/nrg3394

25. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. ArrayExpress - A public repository for microarray gene expression data at the EBI. Nucleic Acids Research. 2003. pp. 68–71. doi:10.1093/nar/gkg091

26. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet. 1999;21: 33–37. doi:10.1038/4462

27. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18: 1509–1517. doi:10.1101/gr.079558.108

28. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (80-). 1995;270: 467–470. doi:10.1126/science.270.5235.467

29. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. NIH Public Access; 2009;10: 57–63. doi:10.1038/nrg2484

30. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial Analysis of Gene Expression. Science (80-). 1995;270: 484–487. doi:10.1126/science.270.5235.484

31. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics. BioMed Central; 2004;5: 18. doi:10.1186/1471-2105-5-18

32. Obayashi T, Nishida K, Kasahara K, Kinoshita K. ATTED-II updates: Condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. Plant Cell Physiol. 2011;52: 213–219. doi:10.1093/pcp/pcq203

33. Hamada K, Hongo K, Suwabe K, Shimizu A, Nagayama T, Abe R, et al. OryzaExpress: An integrated database of gene expression networks and omics annotations in rice. Plant Cell Physiol. 2011;52: 220–229. doi:10.1093/pcp/pcq195

34. Mochida K, Uehara-Yamaguchi Y, Yoshida T, Sakurai T, Shinozaki K. Global landscape of a Co-expressed gene network in barley and its application to gene discovery in triticeae crops. Plant Cell Physiol. 2011;52: 785–803. doi:10.1093/pcp/pcr035

35. Fukushima A, Nishizawa T, Hayakumo M, Hikosaka S, Saito K, Goto E, et al. Exploring Tomato Gene Functions Based on Coexpression Modules Using Graph Clustering and Differential Coexpression Approaches. Plant Physiol. 2012;158: 1487–1502. doi:10.1104/pp.111.188367

36. Obayashi T, Kinoshita K. COXPRESdb: A database to compare gene coexpression in seven model animals. Nucleic Acids Res. 2011;39. doi:10.1093/nar/gkq1147

37. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38. doi:10.1093/nar/gkq537

38. Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, et al. AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. Nucleic Acids Res. 2015;43: D996-1002. doi:10.1093/nar/gku1053

39. Kotera M, Yamanishi Y, Moriya Y, Kanehisa M, Goto S. GENIES: gene network inference engine based on supervised analysis. Nucleic Acids Res. 2012;40: W162-7. doi:10.1093/nar/gks459

40. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M. Interrelating different types of genomic data, from proteome to secretome: ’Oming in on function. Genome Res. 2001;11: 1463–1468. doi:10.1101/gr.207401

41. Alexeyenko A, Sonnhammer ELL. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Res. 2009;19: 1107–16. doi:10.1101/gr.087528.108

42. Gene Ontology Consortium TGO. Gene Ontology Consortium: going forward. Nucleic Acids Res. Oxford University Press; 2015;43: D1049-56. doi:10.1093/nar/gku1179

43. Bhat P, Yang H, Bögre L, Devoto A, Paccanaro A. Computational selection of transcriptomics experiments improves Guilt-by-Association analyses. Provart NJ, editor. PLoS One. Public Library of Science; 2012;7: e39681. doi:10.1371/journal.pone.0039681

44. Stuart JM, Segal E, Koller D, Kim SK. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science (80-). 2003;302: 249–255. doi:10.1126/science.1087447

45. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. 1998;95: 14863–14868. doi:10.1073/pnas.95.25.14863

46. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998;9: 3273–97. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=25624&tool=pmcentrez&rendertype=abstract

47. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res. 2002;12: 37–46. doi:10.1101/gr.205602

48. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res. 2004;14: 1085–94. doi:10.1101/gr.1910904

49. Han X, Yin L, Xue H. Co-expression analysis identifies CRC and AP1 the regulator of Arabidopsis fatty acid biosynthesis. J Integr Plant Biol. 2012;54: 486–99. doi:10.1111/j.1744-7909.2012.01132.x

50. Maeda H, Yoo H, Dudareva N. Prephenate aminotransferase directs plant phenylalanine biosynthesis via arogenate. Nat Chem Biol. 2011;7: 19–21. doi:10.1038/nchembio.485

51. Persson S, Wei H, Milne J, Page GP, Somerville CR. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci

U S A. 2005;102: 8633–8. doi:10.1073/pnas.0503392102

52. Childs KL, Davidson RM, Buell CR. Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS One. 2011;6: e22196. doi:10.1371/journal.pone.0022196

53. Beiki H, Nejati-Javaremi A, Pakdel A, Masoudi-Nejad A, Hu Z-L, Reecy JM. Large-scale gene co-expression network as a source of functional annotation for cattle genes. BMC Genomics. London: BioMed Central; 2016;17: 846. doi:10.1186/s12864-016-3176-2

54. He F, Karve AA, Maslov S, Babst BA. Large-Scale Public Transcriptomic Data Mining Reveals a Tight Connection between the Transport of Nitrogen and Other Transport Processes in Arabidopsis. Front Plant Sci. Frontiers; 2016;7: 1207. doi:10.3389/fpls.2016.01207

55. de la Fuente A. From "differential expression" to "differential networking" - identification of dysfunctional regulatory networks in diseases. Trends in Genetics. 2010. pp. 326–333. doi:10.1016/j.tig.2010.05.001

56. Anglani R, Creanza TM, Liuzzi VC, Piepoli A, Panza A, Andriulli A, et al. Loss of connectivity in cancer co-expression networks. PLoS One. 2014;9. doi:10.1371/journal.pone.0087075

57. Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, Schubert F, et al. A Network of Multiple Regulatory Layers Shapes Gene Expression in Fission Yeast. Mol Cell. 2007;26: 145–155. doi:10.1016/j.molcel.2007.03.002

58. Carlberg C, Molnar F. Mechanisms of gene regulation. Mechanisms of Gene Regulation. 2014. doi:10.1007/978-94-007-7905-1

59. Spangler JB, Ficklin SP, Luo F, Freeling M, Feltus FA. Conserved Non-Coding Regulatory Signatures in Arabidopsis Co-Expressed Gene Modules. PLoS One. 2012;7. doi:10.1371/journal.pone.0045041

60. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet. 2012;13: 840–52. doi:10.1038/nrg3306

61. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. 2011;12: 7–18. doi:10.1038/nrg2905

62. Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, et al. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Rep. 2014;8: 2015–2030. doi:10.1016/j.celrep.2014.08.019

63. Farnham PJ. Insights from genomic profiling of transcription factors. Nat Rev Genet. 2009;10: 605–616. doi:10.1038/nrg2636

64. Yu C-P, Lin J-J, Li W-H. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Sci Rep. 2016;6: 25164. doi:10.1038/srep25164

65. Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, et al. Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2011;108: 14992–7. doi:10.1073/pnas.1103202108

66. Koryachko A, Matthiadis A, Ducoste JJ, Tuck J, Long TA, Williams C. Computational approaches to identify regulators of plant stress response using high-throughput gene expression data. Curr Plant Biol. 2015;3: 20–29. doi:10.1016/j.cpb.2015.04.001

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

68. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S, Rosenblatt F, et al. Machine Learning and Its Applications to Biology. PLoS Comput Biol. Public Library of Science; 2007;3: e116. doi:10.1371/journal.pcbi.0030116

69. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. Nature Research; 2015;16: 321–332. doi:10.1038/nrg3920

70. Li Y, Chen C yu, Kaye AM, Wasserman WW. The identification of cis-regulatory elements: A review from a machine learning perspective. BioSystems. 2015. pp. 6–17. doi:10.1016/j.biosystems.2015.10.002

CHAPTER 2

UTILITY AND LIMITATIONS OF USING GENE EXPRESSION DATA TO IDENTIFY FUNCTIONAL ASSOCIATIONS[1]

[1] The work described in this chapter was published in the following manuscript:

Sahra Uygun, Cheng Peng, Melissa D. Lehti-Shiu, Robert L. Last, Shin-Han Shiu (2016) Utility and limitations of using gene expression data to identify functional associations. PLOS Computational Biology. http://dx.doi.org/10.1371/journal.pcbi.1005244

Contribution: I was involved in conceptualization, data collection and analyses of this project. I wrote the original manuscript and all authors were involved in reviewing and editing of the manuscript.

2.1 Abstract

Gene co-expression has been widely used to hypothesize gene function through guilt-by association. However, it is not clear to what degree co-expression is informative, whether it can be applied to genes involved in different biological processes, and how the type of dataset impacts inferences about gene functions. Here our goal is to assess the utility and limitations of using co-expression as a criterion to recover functional associations between genes. By determining the percentage of gene pairs in a metabolic pathway with significant expression correlation, we found that many genes in the same pathway do not have similar transcript profiles and the choice of dataset, annotation quality, gene function, expression similarity measure, and clustering approach significantly impacts the ability to recover functional associations between genes using *Arabidopsis thaliana* as an example. Some datasets are more informative in capturing coordinated expression profiles and larger data sets are not always better. In addition, to recover the maximum number of known pathways and identify candidate genes with similar functions, it is important to explore rather exhaustively multiple dataset combinations, similarity measures, clustering algorithms and parameters. Finally, we validated the biological relevance of co-expression cluster memberships with an independent phenomics dataset and found that genes that consistently cluster with leucine degradation genes tend to have similar leucine levels in mutants. This study provides a framework for obtaining gene functional associations by maximizing the information that can be obtained from gene expression datasets.

2.2 Introduction

With the ease of sequencing, an ever increasing number of genomes from a wide range of species are available. One major challenge is to ascribe functions to genomic features. For example, while ~70% of *Arabidopsis thaliana* genes have annotated functions [1], only ~40% of these annotations are supported by experimental evidence such as mutant phenotype or biochemical assays [2]. To increase functional information, transcriptome data have been used to develop hypotheses of gene function based on similarity of expression patterns (co-expression) with genes that have known functions [2–4]. The relationship between co-expression and functional correlation was first shown with *Saccharomyces cerevisiae* and human transcriptome data [5–8]. Subsequently, a large number of plant studies used co-expression analysis to infer gene functions [9–17]. For example, the *MYB28* and *MYB29* transcription factors are co-expressed with the glucosinolate pathway genes that they regulate [9]. Similarly, the transcription factors *CRC* and *AP1* co-express with 58 fatty acid biosynthesis genes, and *crc* and *ap1* mutants have altered fatty acid compositions [15]. More broadly, methods based on integration of multiple types of omics datasets were developed to account for different levels of regulation and to improve gene functional inferences [18–21]. In these data integration exercises, transcriptome data remain the most abundant and the most effective in capturing gene functional relationships [2,18]. Thus, analysis of gene expression results can inform hypotheses of plant gene functions.

Despite its utility, there are known computational and biological limitations in using co-expression for gene functional inference, and these usually are not evaluated in co-expression based studies [2]. First, genes with similar expression profiles may not necessarily have related functions [22]. Second, for those genes that do have related functions, transcription patterns may not be coordinated due to post-transcriptional and other levels of regulation [23]. Third, it is also possible that they do in fact co-express, but that the co-expression criteria need to be optimized. For example, using an expression coherence (EC) measure, which is the ratio of the number of co-expressed gene pairs to the total number of gene pairs [24], only 41% of the Gene Ontology Biological Process (GO-BP) terms have higher ECs than expected by chance [25]. The 59% of pathways with low ECs may contain genes that are regulated beyond transcription. Alternatively, a more detailed exploration is required to determine how co-expression should be defined. Consistent with this, in most studies, a fixed threshold of expression similarity is used to identify pairs of co-expressed genes. Depending on the value of this threshold, the degree of co-expression might be over- or underestimated and lead to false positive or negative associations. Therefore, it is necessary to optimize the criteria used to define co-expression to increase the utility of expression data in guilt-by-association studies.

One major parameter that impacts co-expression studies is the type of dataset; it is expected that not all expression profiling experiments will be informative for revealing functional relationships between any given gene pair [26]. Most studies combine multiple datasets for gene function inference [9,25]. One advantage of this approach is the increased statistical power for establishing correlations. Small number of samples might lead to statistically unreliable connections [27]. However, the inclusion of too many samples can result in the loss of information [28], and expression datasets that are directly relevant to the underlying biological processes might be more useful in functional inference. For example, to uncover drought response pathway genes, it would be better to use a more specific, drought stress dataset instead of a collection that includes potentially uninformative experiments [13]. Other factors that impact the effectiveness of co-expression studies include the specific samples used (e.g. stress vs. developmental series), method of data transformation (e.g. fold change vs. absolute expression values), and the procedures and parameters used to define co-expression. A comprehensive study evaluating the above is needed and would be highly informative for future studies that use co-expression as a means for functional inference.

In addition to inferring functional relationships between two genes, co-expression is useful for uncovering groups of genes with related functions (referred to as clusters). Unsupervised learning methods, particularly various clustering algorithms, are among the most common approaches used to identify co-expression clusters [29]. Once the clusters are identified, functional categories such as GO can be used to evaluate what types of genes are over-represented in each cluster, and gene functions can be hypothesized based on cluster membership [30]. Although clustering and enrichment analyses are straightforward, there is no single best method [31] as there are a large numbers of clustering algorithms and the cluster memberships (which genes are in the same cluster) depend on many clustering variables (e.g. algorithm, distance measure and number of clusters). Because differences in parameter choice strongly influence the types of co-expression clusters obtained, it is important to perform clustering with multiple parameters rather than relying on a single method.

In this study, our goal was to maximize the information from co-expression data to improve predictions of functional associations between genes. Specifically, we asked to what extent *A. thaliana* genes are co-expressed in each metabolic pathway. We also explored the features of high EC pathways. Next, we evaluated the influence of dataset on EC for each metabolic pathway, the best practices in using co-expression to identify novel genes that function in a biological process, and the impact of different commonly-used clustering algorithms and parameters on the ability to identify genes that function in the same pathways. Finally, the biological relevance of cluster membership was validated using an independent phenomics dataset. Overall, we demonstrated that optimizing the use of co-expression based approaches requires considerations of the pathway of interest, expression dataset and clustering algorithm.

2.3 Results and discussion

2.3.1 The extent to which genes in pathways have correlated transcript profiles

To evaluate the extent to which genes with similar expression patterns have similar functions, we asked whether genes in the same *A. thaliana* metabolic pathway were co-expressed (see Methods). To address this question, Pearson Correlation Coefficients (PCCs) between genes in each of the 382 *A. thaliana* metabolic pathways in AraCyc were calculated using an expression dataset consisting of 16 different environmental conditions (referred to as the stress dataset [32]) (**Figure 2.1A**). To broadly examine groups of functionally related genes in addition to metabolic pathways, we also calculated PCCs between genes in each of the 1,710 *A. thaliana* Gene Ontology Biological Process (GO-BP). A group of genes in an AraCyc pathway or a GO-BP is referred to as a “functional category”. The median PCC values were <0.1 for ~60% of functional categories, suggesting that many genes in the same pathway have dissimilar transcript profiles under stress conditions. To assess statistical significance and control for false positive expression correlation, the PCC values of pairs of genes in the same functional category were compared to PCC values of random gene pairs (**Figure 2.1A**). The 95th percentile PCC value of random gene pairs (referred to as PCC95) was 0.41 for the stress dataset. In other words, only 5% of random gene pairs have PCC values >0.41. We used PCC95 as the threshold for calling the expression profiles of a gene pair as significantly positively correlated with a 5% false positive rate. Based on this threshold, only 19% of gene pairs within a functional category have significantly correlated expression patterns.

To determine whether some functional categories contain more members with highly correlated expression than others, we adopted the expression coherence (EC) measure, which ranges from 0 to 1 [25]. Here the "pathway EC" is defined as the proportion of pairs of genes in a pathway or GO category that have significantly correlated transcript profiles. Note that the median ECs in *A. thaliana* are only 0.11 for GO-BPs and 0.14 for AraCyc pathways, indicating that 50% of the functional categories have <11-14% gene pairs with significant expression correlation. Consistent with an earlier study [25], we found that genes in functional categories generally have higher ECs than groups consisted of randomly selected genes (Mann-Whitney test, *p* <2.2e-16; **Figure 2.1B**). In particular, 36% of the AraCyc pathways have higher EC values than the 95th percentile of the random EC distribution (**Figure 2.1B**); these are defined as “high EC pathways”. Similarly, 32% of the GO-BPs have higher EC values than the 95th percentile of the random EC distribution (referred to as “high EC GO-BPs”). One explanation for the slightly higher number of high EC pathways than that of high EC GO-BPs may be because metabolism related pathways tend to have a more highly coordinated transcriptional regulation compared to other types of functional categories. Consistent with this notion, GO-BP categories related to metabolism, including metabolic pathways (GO:0008152) and its child terms, have higher median ECs (0.14 and 0.13) compared to signal transduction (GO:0007165, EC=0.11), cell-cycle (GO:0007049, EC=0.10) and response to stress (GO:0006950, EC=0.08) categories. Among metabolic GO-BPs, amino acid metabolism pathways (GO:0006520) have the highest median EC (0.21) among the categories we compared (**Figure S.2.1**). Overall, GO-BPs have lower ECs than AraCyc pathways (Mann-Whitney Test, *p*=2.41e-03; **Figure S.2.1**).

The ECs for functional categories have a very wide range (**Figure 1B**, **Figure S.2.1**). The differences in ECs may be due to technical issues such as functional annotation quality or methodological issues such as the similarity measure used to assess co-expression. The EC differences can also be due to differences in the biological characteristics of pathways, for example, the role of the pathway, presence of common transcriptional regulatory mechanisms, and regulation at levels beyond transcription. Finally, the dataset used to calculate EC could also be a major factor. In the following sections, we assess the factors influencing ECs and identify ways to maximize ECs for functional categories. Considering false positive annotation can have a significant, negative impact in further analyses, we examined features of high EC categories and the impact of multiple factors on ECs by focusing on AraCyc metabolic pathways in the following sections.

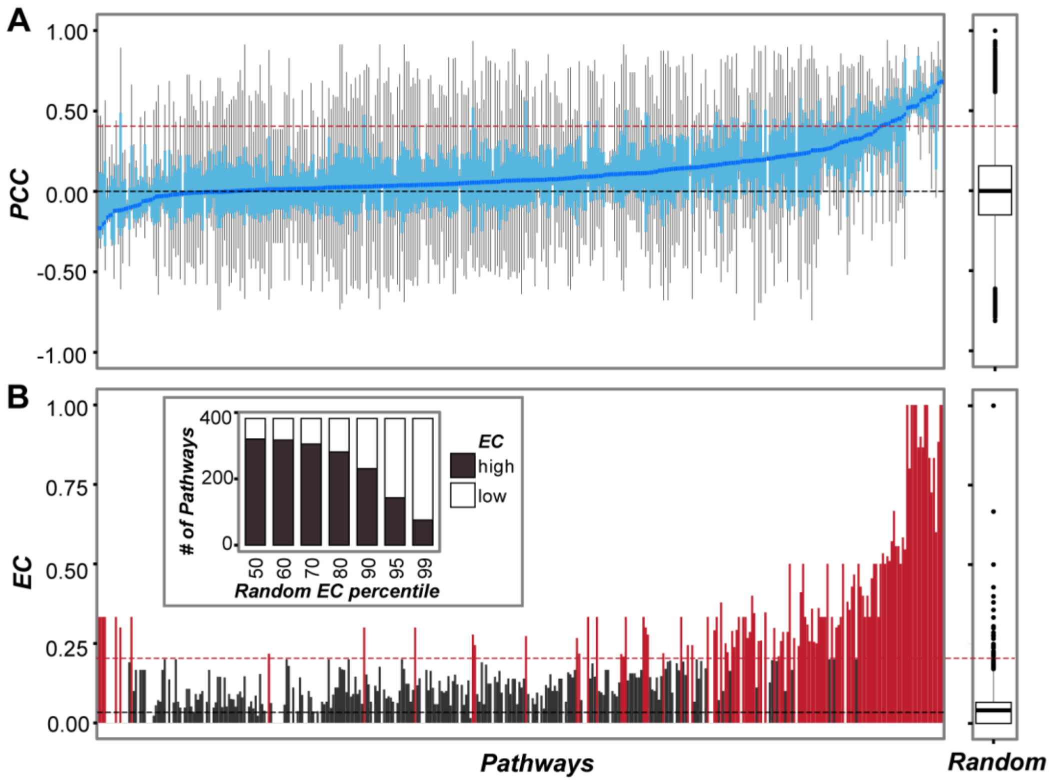


Figure 2.1 Co-expression of *A. thaliana* pathway genes under stress. (A) Boxplots of expression correlations (Pearson's Correlation Coefficient, PCC) between pairs of genes in each *A. thaliana* metabolic pathway (left sub-figure) and random gene pairs (right sub-figure). The pathways are sorted based on median PCC. Light blue boxes: Interquartile range. Blue line: median PCCs. Red dashed line: the 95th percentile PCC value (PCC95=0.41) of the random gene pair PCC distribution. Black dashed line: the median PCC of the random gene pair PCC distribution. **(B)** Bar plot indicating ECs for *A. thaliana* pathways (left sub-figure) and random gene pairs (right sub-figure). The pathways are in the same order as in (A). The insert graph shows the number of pathways that have significantly higher ECs than randomly expected (black) and those that are not significant (white). Different percentile thresholds on the *x*-axis are based on the random EC distribution (right sub-figure). The red dashed line designates the 95th percentile of the random EC distribution.

2.3.2 Influence of annotation on pathway ECs

Computational predictions of gene function without experimental evidence can lead to false assignments to pathways, resulting in lower pathway EC values. This is particularly important because computational annotations in the Plant Metabolic Network are based on sequence similarity only [33,34]. Functional annotations made using sequence similarity based methods are estimated to have an error rate of 49% [35] and high sequence similarity does not necessarily lead to co-expression [36] . To determine whether annotation quality is a major factor influencing pathway EC, we separated pathway genes into those with and without experimental evidence. Consistent with the hypothesis that annotation quality can significantly impact pathway EC, pathways with lower ECs tended to have proportionally fewer genes with experimental evidence (PCC=0.20, *p*=1.53e03; **Figure S.2.2A**). Pathway ECs calculated using genes with experimental evidence were substantially higher (Mann-Whitney test, *p*=5.44e-12, median EC=0.26) than those calculated using genes assigned to pathways solely based on computational predictions (median EC=0.10; **Figure 2.2A**). This is consistent with the hypothesis that some annotations based solely on computational evidence are incorrect.

Although annotation quality influences pathway EC, it explains only ~4% of the variance in the median EC of pathways that include genes assigned based on all evidence (computational and experimental, EC=0.14) and pathways that include genes assigned based on computational evidence (EC=0.10). The small increase in co-expressed genes pairs when including experimental evidence is potentially due to the small fraction of genes that have experimental evidence (5,991 genes considering all evidence, 934 genes considering experimental evidence). Nonetheless, because annotation quality did have a measurable impact, only genes with experimental evidence were included in further analyses.

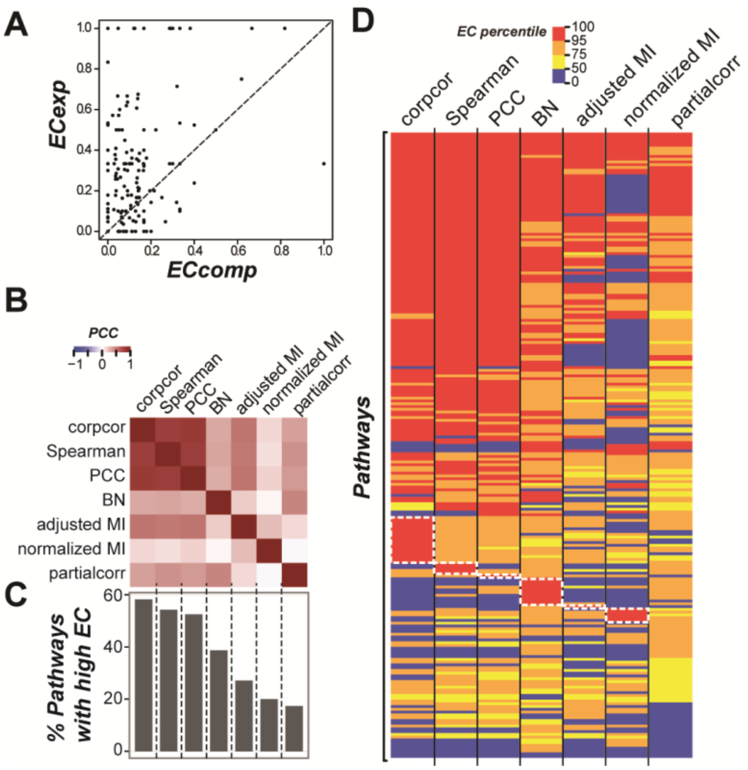


Figure 2.2 Relationship between pathway ECs, annotation quality and similarity measures. (A) Relationship between the EC calculated for pathway genes that are annotated based on experimental evidence (ECexp) and EC calculated for pathway genes that are annotated only computational evidence annotations (ECcomp). The genes used to calculate ECexp and ECcomp do not overlap. Each dot represents one pathway. Dashed line: *y*=*x* line. **(B)** Heatmap of correlations between pathway EC percentiles calculated with: partial correlations estimated with the corpcor method, Spearman's rank correlation coefficient (Spearman), Pearson Correlation Coefficient (PCC), adjusted and normalized Mutual Information (MI), partial correlation calculated with the partialcorr method, and transformed *p*-values of Bayesian Network (BN) **(C)** Percent pathways that have high EC using different similarity measures. **(D)** Heatmap of pathway EC percentiles calculated using different similarity measures. Color represents EC percentiles. White dotted rectangles: high EC pathways that are specific to one measure.

2.3.3 Influence of the similarity measure used to assess EC

In addition to gene annotation quality, similarity measure used to assess gene co-expression could impact pathway EC. Although PCC is among the most widely used similarity measures in co-expression studies, it does not deal with non-linear relationship well as other similarity measures including Spearman's rank correlation coefficient and mutual information (MI). Another consideration is that, all three similarity measures above consider only pairwise correlations, thus higher order correlations due to the influence of the other genes in the network are not considered. To assess the influence of higher order correlation, we also evaluated two approaches: (1) partial correlation, where the correlation between genes was calculated after controlling the effects of other genes and (2) a graph model-based approach such as Bayesian Network (BN) where the strength of connection of a gene pair is determined by considering all genes in a network. To assess the impact of potential non-collinearity and higher order correlations, we first calculated pathway ECs with seven different similarity measures including PCC, Spearman's rank, two partial correlation methods (corpcor and partialcorr), adjusted and normalized MI, and transformed *p*-value of arc strength in a pathway BN (see **2.5 Methods**). To assess the statistical significance of EC values and control for false positive ECs, EC values were calculated with randomly chosen gene pairs for each pathway size and for each similarity measure. Thus, for each pathway size and measure, a random EC distribution is available and used to determine the percentile value of a pathway EC (referred to as "EC percentile"). Thus, a high EC percentile indicates reduced probability that the observed pathway EC is spurious.

First, we asked if the pathway EC percentiles are correlated among different measures (**Figure 2.2B**). For example, PCC were significantly positively correlated with, in order of diminishing degrees of correlations, corpcor (PCC=0.80, *p*=2.35e-50), Spearman's rank coefficient (PCC=0.78, *p*=1.67e-46), adjusted MI (PCC=0.53, *p*=5.23e-18), partialcorr (PCC=0.39, *p*=1.97e-09), BN (PCC=0.30, *p*=5.35e-06), and normalized MI (PCC=0.17, *p*=1.23e-02). Given the degrees of correlations in EC percentiles differ widely between measures, the similarity measures have significant impact on pathway ECs. Consistent with this notion, the number of pathways with ECs that are significantly higher than randomly expected (high EC pathways, >95th percentile of the random EC distribution) vary widely depending on the similarity measure (**Figure 2.2C**). Among the measures, corpcor, PCC and Spearman's rank allowed the highest numbers of high EC pathways to be identified. This finding is consistent with the finding of a recent study examining PCC, Spearman's rank coefficient, MI, and other similarity measures [27]. Only five of the pathways have high ECs consistently regardless of similarity measures (**Figure 2.2D**). Importantly, consistent the idea that non-linearity and higher order correlations can be important, the ECs of some pathways are only significant if a particular similarity measure is used (white box, **Figure 2.2D**). Notably, 17 and 10 pathways have high ECs only when the corpcor method and the BN-based measure were used, respectively (**Figure 2.2D**), illustrating the importance of higher order correlations. In addition, different methods of calculating partial correlations led to significant differences in high EC pathway recovery. As the corpcor method was optimized for genomic data analysis [37], it is not surprising that the results from corpcor is more informative. For further analyses, we continue with PCC as the measure the calculate gene co-expression as it is one of the widely-used similarity measure and, along with Spearman's rank and corpcor, uncover the highest numbers of high EC pathways.

2.3.4 Influence of biological factors on pathway EC

Next, we explored biological factors that may influence pathway EC, including pathway size (the number of genes assigned to the pathway), subcellular location, pathway gene function, and evidence of co-regulation. We hypothesized that a pathway with a larger number of genes might have relatively more complicated modes of regulation beyond transcription, leading to low pathway ECs. In addition, gene products with similar functions tend to be co-localized and may be coordinately regulated [38], as in the case for photosynthesis and other chloroplast-related pathways [39]. However, pathway gene number was not significantly correlated with pathway EC (PCC=-0.03, *p*=0.67; **Figure 2.3A**), and pathway gene product subcellular location was not associated with pathway EC (**Figure 2.3B**).

To assess whether the general biological functions of a pathway contribute to differences in EC between pathways, we compared EC between five general pathway categories including generation of precursor metabolites and energy, biosynthesis, degradation. However, the significance of enrichment of these general categories was only marginal (Mann-Whitney Test, *p*=0.05; **Figure 2.3C**). Interestingly, although the expression of gene pairs in the general category of generation of precursor metabolites and energy is not always significantly coherent, the specific pathways within—photosynthesis light reactions, chlorophyllide a biosynthesis I and aerobic respiration—had significantly higher ECs compared to random pathways (99th percentile of pathway EC distribution). This finding suggests that EC, and more generally co-expression, is more relevant to more detailed levels of the functional classification hierarchy.

Transcriptional regulation is another major factor that could influence pathway EC. Genes that are co-regulated could have similar transcript profiles, and the differences in the degree of co-regulation may explain differences in pathway EC. To determine the extent of co-regulation, we asked how the presence of *cis*-regulatory elements differs among pathways. It is expected that pathway genes with similar sets of *cis*-elements in their promoters would have similar expression patterns and thus contribute to high pathway EC. We mapped 349 transcription factor binding motifs [40] to the promoters of all *A. thaliana* genes, and identified motifs that were over-represented in the promoters of pathway genes taking each pathway separately and comparing to all other genes. A total of 40 overrepresented motifs were found for 17 pathways. However, there was no significant difference in EC between pathways with and without overrepresented motif sites (Mann-Whitney Test, *p*=0.66; **Figure S.2.2B**). This was surprising given that the 349 motif dataset spans essentially all known *A. thaliana* transcription factor families, and transcription factors from the same family tend to have similar binding motifs [40]. Thus, the reason why high EC pathway genes do not necessarily have more shared motifs (**Figure S.2.2B**) is not simply due to unknown transcription factor binding sites. This finding can also be due to complex interactions between binding sites, nucleosome positioning and other DNA properties [41]. We also evaluated post-transcriptional regulation by miRNA, but did not find a significant difference in EC between pathways with miRNA target genes and those that did not (Mann-Whitney Test, *p*=0.31; **Figure S.2.2C**). Given the dearth of genome-wide post-transcriptional and other levels of regulatory data in plants, it remains to be resolved if post-transcriptional regulation contributes to a lower pathway EC.

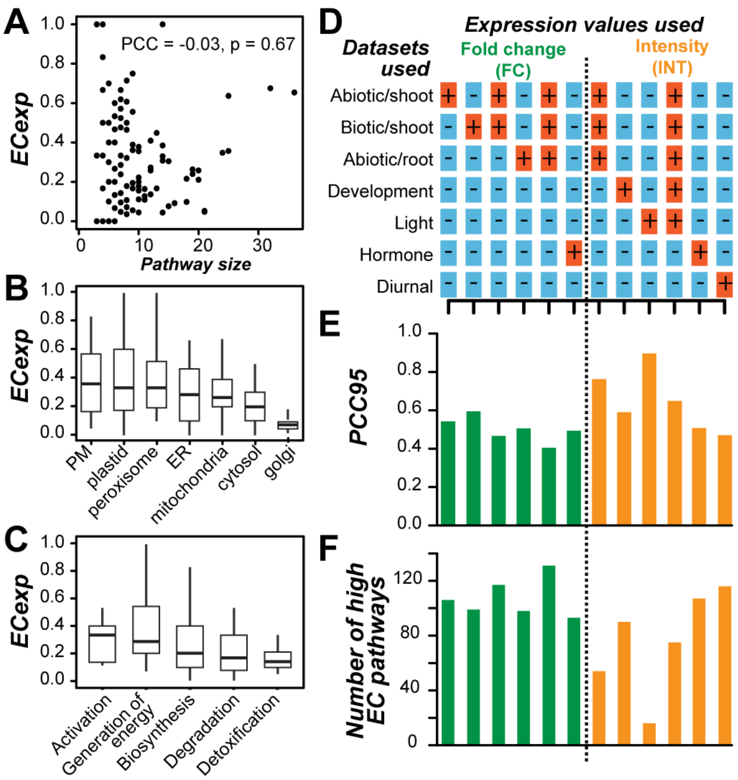


Figure 2.3 Impact of pathway size and other factors on EC. (A) Relationship between ECexp of a pathway and pathway size (the number of genes assigned to a pathway). **(B)** ECexp value distribution for pathway genes with products that have subcellular location annotations. PM: Plasma membrane **(C)** ECexp value distribution for different pathway classes (general pathway categories). **(D)** Datasets used to determine pathway ECs. A "+" indicates that the dataset in question was used (either individually or in combination) for the analyses depicted by bar graphs in (E) and (F). The columns in (D) correspond to those in (E) and (F). **(E)** The 95th percentile PCC values (PCC95) in the null distributions for each dataset or combination of datasets. PCC95 of combined datasets (stress fold change and light (L)+stress (S)+development (D) absolute intensity) are labeled in the bar plot **(F)** Number of pathways with high EC for each dataset and or combination of datasets. Green: fold change values were used to calculate ECs. Orange: absolute intensity values were used for calculating ECs.

2.3.5 Impact of datasets used to evaluate pathway EC

Among the factors studied ─ the size of the pathway, subcellular location, functions of pathway genes, and evidence of shared transcription factor binding sites ─ none significantly impact pathway EC. We next asked whether the expression dataset has a major impact on whether the EC for a pathway is high or low. The analyses described so far were performed using an environmental stress dataset consisting of 112 experiments including biotic and abiotic stress treatments in shoot and root [32]. Low pathway EC values could reflect the fact that pathways are only relevant to one type of stress (biotic or abiotic) and a large compiled dataset fails to capture the underlying patterns of co-expression. To address this possibility, we first calculated the random gene pair correlations for three subsets of the environmental stress gene expression dataset: shoot abiotic, shoot biotic, and root abiotic. PCC95 values were higher for subsets (PCC95=0.51 - 0.60) of the stress dataset than for the entire dataset (PCC95=0.41; **Figure 2.3D and E**), indicating that the difference in gene expression between experiments within a dataset, i.e. data heterogeneity, was lower when the samples were divided into biologically relevant subsets. Consistent with this, the average sample correlation within each of the shoot biotic, shoot abiotic, and root abiotic subsets is higher (0.15, 0.19, and 0.46, respectively) than the entire environmental stress dataset (0.13, Mann-Whitney Test using all pairwise sample PCCs, *p*=8.73e-26, 2.10e-05, 1.93e-144). Due to the impact of data heterogeneity, fewer high EC pathways tend to be recovered from individual stress datasets compared to combined datasets (**Figure 2.3F**).

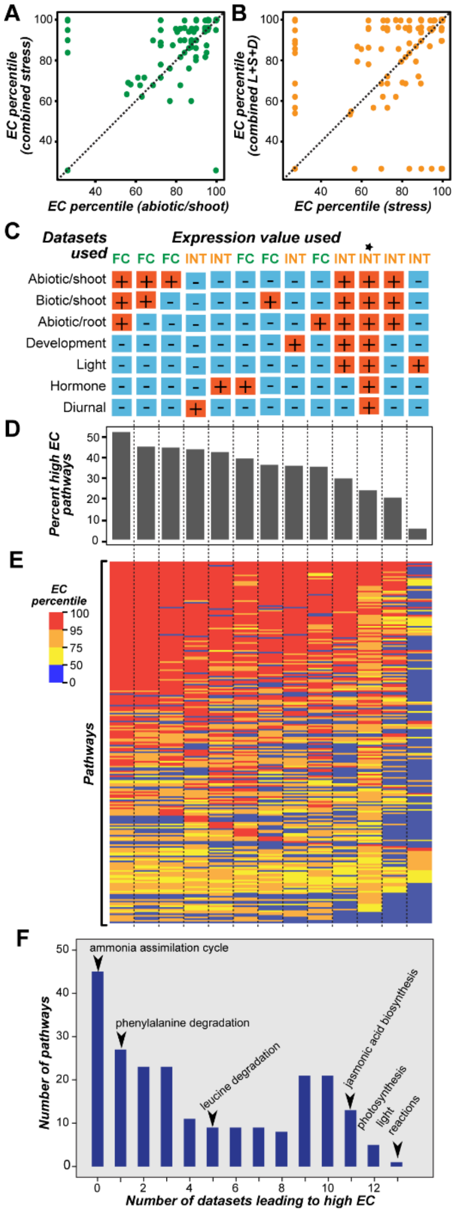
To test whether these findings are specific to the environmental stress data, an additional four expression datasets were analyzed (development, light, hormone, and diurnal; **Figure 2.3D**). We found that the threshold PCC95 values of these datasets were significantly negatively correlated with the number of high EC pathways (PCC=0.97, *p*=1.10e-08). Thus, because PCC95 is negatively correlated with data heterogeneity (as discussed in the previous section), higher data heterogeneity likely allows more co-expressed pathway genes to be recovered. Data heterogeneity can be influenced by which datasets are combined and how the expression data are processed and transformed. Combining datasets tends to increase data heterogeneity and thus leads to a better recovery of pathway genes based on co-expression (**Figure 2.3F**). Dataset processing also has an effect on data heterogeneity. For example, datasets that were processed to obtain fold change values had a substantially lower PCC95 (median PCC95 of fold change datasets=0.41; **Figure 2.3E**) than that of the absolute intensity dataset (median PCC95 of intensity datasets=0.76; **Figure 2.3E**), although this was not true for the hormone dataset (**Figure 2.3E**). Taken together, these results reveal that dataset transformation approaches and nature of the expression dataset impact the threshold for defining significant co-expression and thus significantly shapes pathway EC.

2.3.6 Influence of individual vs. combined stress datasets on pathway EC

A wide range (5%-53%) of pathways have significantly high ECs depending on the dataset used (**Figure 2.3F**). This pattern led us to question whether some datasets were more informative than others in recovering specific pathways. To assess this, pathway EC percentiles were calculated for each dataset separately (**Figure S.2.3A**). Note that for each expression dataset analyzed, we picked half a million pairs of randomly chosen genes from a total of ~22,000 to establish background correlations and selecting the correlation threshold at the 95th percentile of the random correlation distribution. Because dataset heterogeneity influenced the threshold values used to determine gene co-expression (**Figure 2.3E**), we first asked whether larger, combined stress datasets were more informative (i.e. had higher pathway EC percentiles) compared to smaller, individual datasets (**Figure 2.4A**; **Figure S.2.3B and C**). The combined stress dataset had a higher median EC percentile (95.6) compared to the individual datasets (89.5-89.9). For example, the monoterpene biosynthetic pathway had an EC percentile of 99.6 based on the combined stress dataset, but the values ranged from 26.3 to 89.9 for individual datasets.

By contrast, in >14% of the pathways, the EC percentiles determined with the individual datasets were higher than those based on the combined dataset (**Figure 2.4A**; **Figure S.2.3B and C**). For example, the lipid dependent phytate biosynthesis I pathway had an EC percentile of 99.5 when the root abiotic stress dataset was used compared with EC percentiles<27 for all other individual and combined datasets. Another example is the cuticular wax biosynthetic pathway, which had an EC percentile of 99.7 calculated from the shoot abiotic stress data, but had EC percentiles of 26.4 and 26.6 when root abiotic and shoot biotic stress datasets were used, respectively. This is consistent with the role of cuticular wax in protecting the shoot from drought and other stresses [42,43] and the co-regulation of its biosynthetic genes [44]. Similarly, indole-3-acetic acid (IAA) degradation genes have EC percentiles of 99.9 and 26.3 using root and shoot abiotic stress datasets, respectively, consistent with the finding that IAA degradation products have been mainly detected in roots [45,46].

These findings lead to the conclusion that EC among genes in the same pathway is strongly influenced by whether individual or combined stress datasets are used, particularly if the pathway in question is biologically relevant to the experimental conditions of the dataset. Thus, it is important to test multiple individual and combined datasets for finding the optimal EC for a pathway. It should be noted that, while high EC pathways can be recovered with individual datasets, the smaller numbers of samples in individual dataset have less power in detecting co-expression. This is because we have included randomized background information for different sized datasets in calculating threshold pairwise similarities for determining EC and in calculating threshold EC values for identifying pathways with significantly high ECs. A smaller dataset where spurious correlations are expected will have a correspondingly higher threshold because the correlations between randomized gene pairs will be higher.



Datasets used	Expression value used												
	FC	FC	FC	INT	INT	FC	FC	INT	FC	INT	★ INT	INT	INT
Abiotic/shoot	+	+	+	-	-	-	-	-	-	+	+	+	-
Biotic/shoot	+	+	-	-	-	-	+	-	-	+	+	+	-
Abiotic/root	+	-	-	-	-	-	-	-	+	+	+	+	-
Development	-	-	-	-	-	-	-	+	-	+	+	-	-
Light	-	-	-	-	-	-	-	-	-	+	+	-	+
Hormone	-	-	-	-	+	+	-	-	-	-	+	-	-
Diurnal	-	-	-	+	-	-	-	-	-	-	+	-	-

Figure 2.4 Impact of datasets on pathway EC Percentile. (A) Relationship between pathway EC percentiles calculated using the combined stress gene expression dataset and those calculated based on individual stress dataset, abiotic/shoot. **(B)** Relationship between pathway EC percentiles calculated using the light, development, and stress combined dataset and those calculated based on individual dataset, stress. In (A) and (B) the dashed line represents *y* =*x*, and each dot represents a pathway. **(C)** Individual and combinations of datasets used to determine pathway EC Percentiles.

Figure 2.4 (cont'd)

*: NASCArray consisting of all the datasets listed here as well as additional datasets (~700 samples). The columns in (C) correspond to those in (D) and (E). **(D)** Bar plot of percent high EC pathways using different expression datasets **(E)** Heat map of pathway EC percentiles from 13 gene expression datasets. Dark red: EC percentiles≥ 95. Orange: 95 > EC percentiles < 75. Yellow: 75 > EC percentiles <50, Blue: 50 > EC percentiles < 0 **(F)** Histogram of the numbers of datasets leading high EC values for each pathway. Example pathways are labeled with an arrow.

2.3.7 Robustness in recovering pathway genes when using different datasets

To determine whether the conclusion that EC is strongly influenced by stress (S) datasets is generalizable to non-stress ones, we further increased the dataset size by including light (L) and developmental series (D). We found that when using dataset L, S, D, and combined (L+S+D) datasets, 12, 46, 81, and 96 pathways had significantly higher than expected EC, respectively. Although the combined dataset was the best for uncovering more pathways, the EC percentiles were higher for some pathways when individual datasets were used (**Figure 2.4B**; **Figure S.2.3D and E**). Two interesting examples are the *trans*-zeatin biosynthesis and the iron reduction/absorption pathways. These pathways only had significantly high EC when using the light dataset. Fluctuations in light conditions can alter the expression of *trans*-zeatin biosynthesis genes [47]. In addition, iron is a central component of chlorophyll. One iron reduction gene, FRO6, contains multiple light-responsive elements, and another, iron reduction gene FRO7, has an expression pattern similar to FRO6 [48,49]. Consistent with our discussion on the impact of individual and combined datasets in the previous section, these findings indicate that dataset choice impacts the optimal recovery of pathway genes. Next, we asked how data transformation impacts pathway EC percentile. The EC percentiles determined from fold change and absolute intensity were significantly positively correlated for the stress (PCC=0.38, *p*=4.90e-9) and hormone (PCC=0.57, *p*=1.17e-20; **Figure S.2.3F and G**) datasets. Despite these significant correlations, data transformation still resulted in a >50 percentile difference in EC for 27% and 12% of pathways using stress and hormone datasets respectively.

Based on our results, it is important to test datasets according to the pathway of interest, but do more expression data samples necessarily lead to better pathway recovery? To answer this question, we compared pathway EC percentiles across 12 individual and combined datasets (**Figure 2.4C**). We found that stress dataset had the highest percentage of high EC pathways (53%) recovery rate among larger, combined datasets analyzed (**Figure 2.4C**). To further assess whether using a much more inclusive, more conditionally independent dataset compared to the 12 datasets we used, would increase the recovery rate of high EC pathways, we analyzed NASCArrays dataset with >700 samples [50]. We found that 24% of the pathways had high EC with the NASCArray dataset. This recovery rate was lower compared to a much smaller dataset such as the stress set, where 53% of pathways had high ECs (**Figure 2.4D**). Thus more is not necessarily better. This is because the overlap in within and between pathway expression correlations was larger when NASCArray dataset was used compared to stress dataset (**Figure S.2.4A and B**), indicating that it was harder to distinguish within and between pathway gene pairs using the NASCArray data.

Next, we asked if some pathways have significantly high EC regardless of the dataset used (i.e. are robust). Among pathways, 180 had significantly high EC in >1 datasets (**Figure 2.4E**), but photosynthesis light reactions was the only pathway that had significantly high EC in all datasets. This is consistent with earlier findings that light reaction genes are tightly co-regulated

[51]. In addition to photosynthesis light reactions, jasmonic acid biosynthesis, aliphatic glucosinolate biosynthesis side chain elongation cycle, fatty acid elongation, palmitate biosynthesis II and chlorophyll a degradation II were also among the most robust pathways in terms of EC.

On the other end of the spectrum, 15% of the 179 pathways with significant EC had significantly high EC in only one dataset (e.g. phenylalanine degradation; **Figure 2.4F**), further indicating the importance of dataset selection for co-expression associations with unknown genes. In addition, 21% of the pathways (e.g. ammonia assimilation cycle; **Figure 2.4F**) did not have significant EC regardless of the dataset used; indicating that additional datasets may be required and/or these pathways are mainly regulated at levels beyond transcription. Given that many pathways had significant EC when a particular dataset was used, we asked how many individual datasets are required to recover the 180 pathways with significant ECs. Interestingly, when datasets are included one at a time, the number of pathways with significantly high EC initially increased but appeared to be saturated after the addition of 11 datasets (**Figure S.2.3H**).

Taken together, although genes within pathways can have similar expression patterns, this similarity is best recovered after experimenting with a number of different individual and combinations of datasets as well as with data transformations. In addition, although data heterogeneity increases the number of pathway genes that can be recovered, combining datasets is not necessarily the best approach for all pathways. Comparing to 5-53% high EC pathways that can be discovered when datasets are used individually, combining the analysis results of the individual datasets led to the finding that 80% pathways have high ECs.

2.3.8 Clusters as predictive units of pathways

Clustering genes based on similar expression profiles is commonly performed to find genes that are functionally related [52]. In the best-case scenario, most of the genes in a pathway would be in the same cluster, and the remaining genes in the cluster could be tested to see if they have functions similar to the pathway genes. To evaluate the extent clustering would give us this best case scenario, we first employed one of the most widely used clustering algorithms, *k*-means, to group ~22,000 genes in the stress gene expression dataset. To determine the optimal *k*, there are multiple proposed statistical methods including Bayesian Information Criterion (BIC) [53], gap statistic based on the elbow plot [54], and silhouette score [55]. Although these measures have been successfully implemented in simulated datasets where the grouping is apparent [56], there is no best method in determining the number of natural groups of the high-throughput genomics data and often researchers have to try multiple number of clustering results [57,58]. In our initial analysis, we used elbow plot to define *k*. We computed within cluster sum-of squares for a range of *k* values starting from 5 clusters and going up to 2000 (**Figure S.2.5A**). Even though there is no clear elbow point, the decrease in the within sum of squares was apparent when *k*=100 which was used for *k*-means clustering. Once the 100 clusters were obtained, over-representation analysis was used to assess how well pathway and cluster membership coincide and an over-representation score was defined (see **2.5 Methods**). Clusters with significant over-representation scores (q <0.05) were analyzed further (**Figure 2.5**). Our expectation for an ideal clustering result was a low q-value (~0). Only 30% of the pathways were found to be over-epresented in >1 cluster, and 38% pathways had an over-representation score < 2 ($0.01 < q < 0.05$).

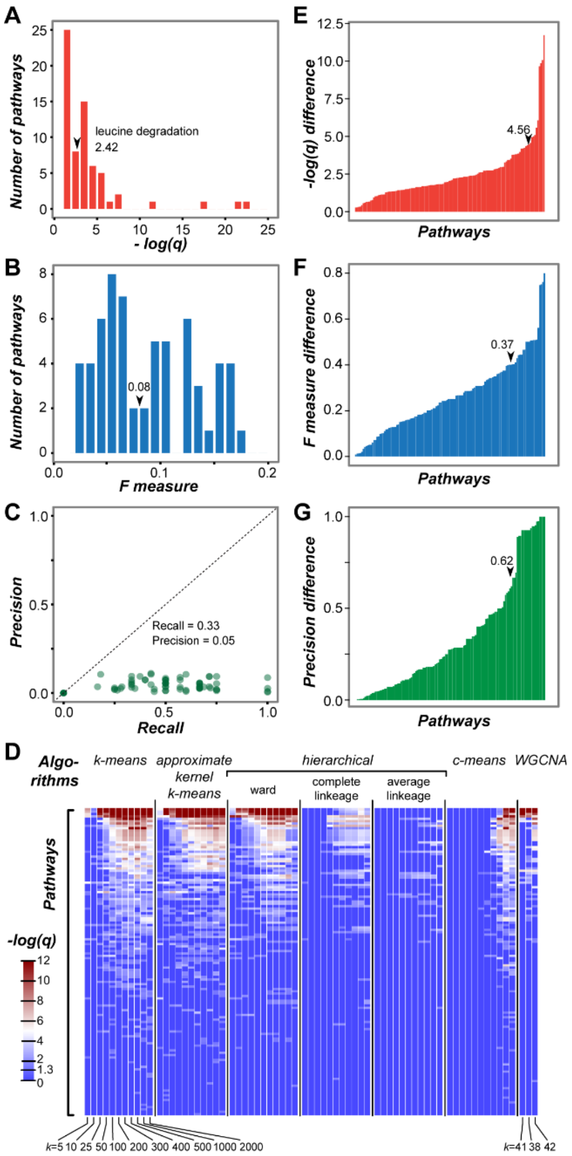


Figure 2.5 Performance of clusters in predicting pathways. (A) Histogram of the maximum scores (-log(*q*)) for over-representation of pathways within clusters. **(B)** Histogram of the maximum F measures for prediction of pathway membership based on cluster membership. **(C)** Relationship between precision and recall for clusters. In (A-C), clusters were generated using k-means with k=100. **(D)** Heat map of over-representation scores obtained from different individual and combined clustering algorithms (top) and cluster numbers (bottom) Color represents over-representation scores (-log(*q*)) from 0 to 12.

Figure 2.5 (cont'd)

Scores less than 1.3 are indicated by dark blue. Scores more than 1.3 are represented by a spectrum of light blue to red. Pathways in the heat map are sorted based on the number of times that they are over-represented in the clusters, high to low. **(E)** Bar plot showing the difference between overall maximum over-representation score ─ the highest score from any single cluster ─ and the over-representation score from clusters generated using *k*-means, *k*=100 for each pathway. **(F)** Bar plot showing the difference between the overall maximum F measure ─ the highest score from any single cluster ─ and the F measure from clusters generated using k-means, k=100 for each pathway. **(G)** Bar plot showing the difference between maximum Precision ─ the highest score from any single cluster ─ and the Precision from clusters generated using k-means, k=100 from each pathway. Arrow: performance values for the leucine degradation pathway

As significance alone does not tell us to what extent each cluster is informative in finding additional genes associated with the pathway of interest, we evaluated each clustering result as a prediction problem, where a gene's membership in a cluster is used to predict its membership in a particular pathway. The performance of the clustering results was evaluated using the F measure, which is the harmonic mean of Precision and Recall. Here precision is the proportion of the number of genes that overlap between a cluster and a pathway to the number of genes in the cluster. Recall is the proportion of the number of genes that overlap between a pathway and a cluster to the number of genes in the pathway. F measures can range from 0 and 1 and higher F measures suggest that both Precision and Recall are high. Precision, Recall and F measures were calculated for every pathway-cluster combination when there was a significant enrichment (q <0.05; **Figure 2.5B and**

C). We expected high Precision (~1) for the most informative clusters, but the highest precision among cluster-pathway combinations was 0.11. In one cluster, 11% of the genes belong to the "glucosinolate biosynthesis from the tryptophan pathway". The same cluster also yielded the highest F measure (0.18). This result suggests that there is a need to improve this clustering result, potentially by using different clustering algorithms and parameters that is explored further in the next section.

2.3.9 Impact of clustering algorithms and parameters on the identification of pathway genes

In the analyses described so far, we used only one clustering algorithm (*k*-means) and fixed parameters (Euclidean distance, *k*=100). Next, we assessed how additional clustering algorithms and clustering parameters (number of clusters defined, distance measure, and number of runs) impact the identification of co-expressed gene clusters and how this in turn impacts the identification of genes with similar functions. Five algorithms were applied to the stress expression dataset using different parameters including number of clusters (*k*), consistency among runs, and other algorithm-specific parameters, to obtain 366 different clustering results. Although some of the algorithms (*k*-means, approximate kernel *k*-means, *c*-means) often yield local optima instead of an overall best result, clustering runs with the same algorithm and parameters gave very similar results (average PCC among 10 runs=0.8 - 1.0). Therefore, only the maximum over-representation score from 10 runs is shown (**Figure 2.5D**).

We found that the choice of *k* is important; regardless of the algorithm, smaller *k* values resulted in low over-representation scores (**Figure 2.5D**) and a smaller number of pathways over-represented among clusters (**Figure S.2.5B**). This is likely due to the fact that smaller *k* values lead to larger sized clusters that contain genes from multiple pathways. We also found that the number of members in a cluster that overlap with members of a pathway differs depending on the algorithm used; *k*-means was the best performing algorithm, followed by approximate kernel *k*-means and hierarchical clustering with the Ward algorithm (**Figure S.2.5B**). Overall, with all clustering methods combined, we were able to recover 131 pathways out of 225 (64 more pathways than when only *k*-means, *k*=100 was used). In contrast, 95 out of 225 pathways were not over represented in any of the clusters, and 22 pathways were only over-represented in one algorithm-parameter combination (**Figure S.2.5C**). Taken together, the clustering approach is not deterministic; the parameters used influence co-expression associations. Therefore, it is important to evaluate multiple algorithms and parameters to recover pathways of interest.

Multiple algorithm-parameter combinations were examined (e.g. an example combination: *k*-means, *k*=100), to quantitatively assess the degree of improvement in performance measures. First, clusters from 69 algorithm-parameter combinations were generated (**Figure 2.5D**). For each pathway, we asked what the maximum over-representation score was among the clusters from all combinations. This maximum score was then compared to the over-representation score of clustering results from our standard method discussed above (*k*-means, *k*=100; **Figure 2.5E**). We found that the over-representation scores of the best clusters were increased by an average of 1.40 (25-fold better *q*-value) compared to the score when only one algorithm/parameter was used. We also evaluated clustering performance using F measure (improved by an average of 0.15; **Figure 2.5F**) and Precision (improved by an average of 0.20; **Figure 2.5G**). These results reinforce the importance of considering multiple algorithms and parameters to maximize pathway-cluster overlap. Furthermore, for algorithms requiring a predefined *k*, the *k* value may be different depending on the pathway one would like to recover and it is necessary to try out multiple values for the best results. Thus, selecting a presumably optimal *k* may yield a more natural grouping of the entire dataset but at the expense of uncovering clusters representing individual pathways.

We should emphasize that, although considering multiple clustering parameters allow recovery of 93 pathways, there are still 96 pathways that were not recovered by the five algorithms used in this study (**Figure S.2.6**). This may be because genes in these pathways do not have highly coordinated expression patterns and have low pathway ECs. Consistent with this interpretation, high EC pathways tend to be recovered by clustering compared to low EC ones (Fishers exact test, *p*=4.56E-12; **Figure S.2.6**). We should also emphasize that the scores used to assess the clustering performance ignore the possibility that some genes in the clusters will be novel pathway components. The presence of these genes reduces the over-representation score, precision, and F-measure. These novel pathway component genes are prime candidates for further functional characterization using genetic or biochemical analysis.

2.3.10 Using leucine degradation gene phenomics data to validate co-expression associations

We established that the degree of gene co-expression in some pathways is influenced by dataset and data transformation and that it is important to use multiple algorithms and parameters when identifying clusters based on co-expression. To demonstrate that novel pathway components can in fact be recovered as a result, we used phenomics data to validate novel gene components of the leucine degradation pathway [59,60]. We chose to focus on leucine degradation because it is among the most over-represented pathways in co-expression clusters (**Figure 2.5**), and many components of the leucine degradation network remain to be discovered in plants [61,62]. Eighteen novel genes that were not annotated to leucine degradation in the AraCyc database are consistently found in clusters (≥10 clustering results) that are over-represented with 12 annotated leucine degradation genes. Among these genes, AT1G55510, a branched-chain alpha-keto acid decarboxylase E1 beta subunit, was recently shown to be involved in leucine degradation [61] but has not yet been annotated as such. The fact that AT1G55510 is consistently found in the same clusters as leucine pathway genes prompted us to examine the rest of the genes that cluster with leucine degradation genes for involvement in leucine degradation.

We hypothesized that previously unknown associations deduced from co-expression clusters could be verified based on their mutant phenotype data. To test this hypothesis, we used a published phenomics dataset that includes free seed leucine levels for mutants in more than 5,000 genes [60] (**Figure 2.6A**). The free leucine levels (nmol/g fresh weight) of leucine degradation gene mutants are expected to be more similar to genes within the same cluster than to wild type plants or randomly chosen mutants. As expected, the leucine degradation enzyme genes had higher leucine levels than mutants in random genes and wild-type plants (*p*=0.05 and 0.04 respectively; **Figure 2.6B**). Next we evaluated the clusters that were over-represented with leucine degradation genes by calculating the log ratio between the proportion of leucine degradation genes in a cluster to the proportion of non-leucine degradation genes in the same cluster. Note that, as *k* increases, the log ratio tends to increase (**Figure 2.6C)**. This trend is potentially due to increased statistical power to identify over-representation in smaller sized clusters. Among these clusters, hierarchical clustering with the Ward algorithm (*k*=100 and *k*=200) and approximate kernel *k*-means (*k*=50, *k*=400 and *k*=500) yielded clusters that had genes (**Figure 2.6D**) whose leucine levels were significantly higher than the wild type measurements (*p*=0.01-0.05; **Figure 6E**). Thus, some genes in those co-expressed clusters are likely involved in leucine degradation. Nonetheless, the differences in leucine levels between mutants of genes in the cluster and wild type plants were small (**Figure 2.6E**). This may be due to the fact that some co-expressed genes are false positives.

However, some known leucine degradation pathway gene mutants also do not have dramatic differences in leucine level compared to wild type (**Figure 2.6B**) and this may also explain the small effect size. We next asked whether a gene that consistently clusters with leucine degradation genes ─ regardless of the algorithm and parameters used ─ tends to be a better pathway gene candidate than one that does not. Mutants in genes that were retrieved from three separate clustering results had significantly higher leucine levels than mutants in random genes and wild-type plants (*p*=0.03 and 2.60e-3 respectively; **Figure 2.6F**), indicating that consistency may serve as a criterion to increase confidence in candidate genes.

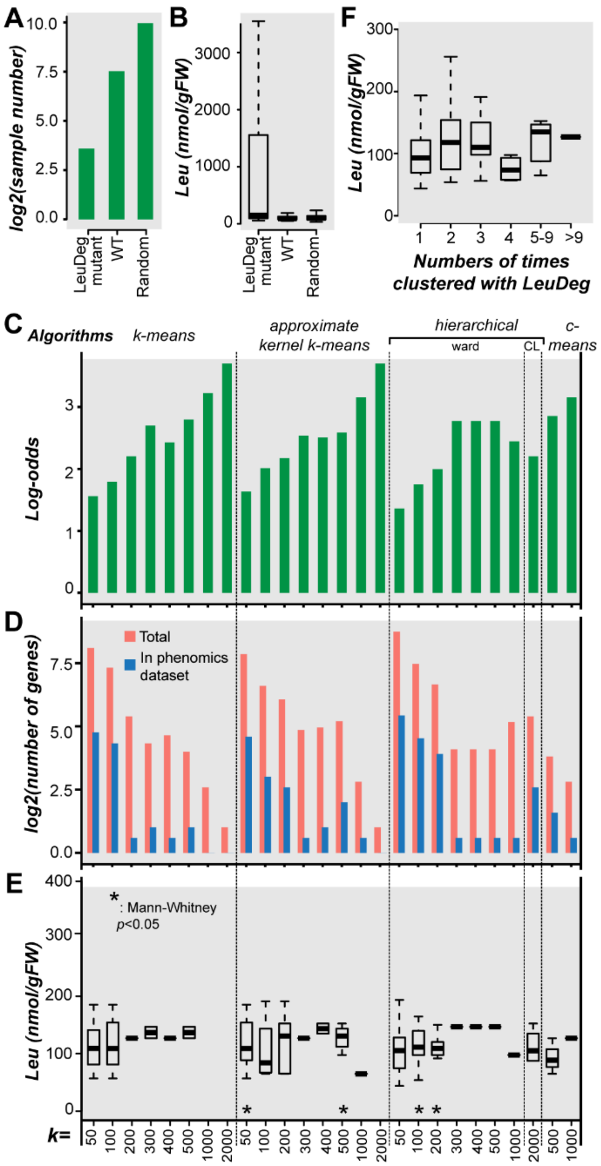


Figure 2.6 Assessing the validity of co-expression associations with leucine measurement data. (A) Log2 of the number leucine degradation (LeuDeg) gene mutants, random gene mutants, and WT control plants that were included in the analysis of leucine levels. **(B)** The absolute leucine levels (nM/gFW) in the same three types of genetic background as in (A). **(C)** Log-odds values (log ratio between the proportion of leucine degradation genes in a cluster to the proportion of non-leucine degradation genes in the same cluster) of clusters that are enriched in leucine degradation genes identified using different algorithm-size parameter combinations.

Figure 2.6 (cont'd)

(D) Log$_2$ of the number of genes that cluster with leucine degradation pathway genes (over-representation score >1.3) for each algorithm-size parameter combination. **(E)** Box plot showing the absolute seed leucine levels (nM/gFW) in plants with T-DNA insertions in genes clustered with leucine degradation pathway genes (enrichment score >1.3) for each algorithm-cluster size parameter combination. *: the groups of genes where mutant leucine levels are significantly greater than in wild type. **(F)** The absolute seed leucine levels (nM/gFW) of T-DNA insertion mutants of genes that cluster with leucine degradation genes. Binning in *x*-axis depends on the number of times that each gene clusters with leucine degradation genes considering all clustering results.

2.4 Conclusion

A large number of high-throughput omics data are accumulating. Of these, transcriptome data are the most abundant, covering multiple tissues and conditions, and have been widely used to generate hypotheses about gene functions. Since almost the first microarray studies, researchers have used the guilt-by-association approach to make useful predictions about gene functions. This approach is based on the hypothesis that genes encoding proteins of shared function are more likely to have common features such as gene expression patterns. Here we show that even though this approach is useful, there are many limitations to co-expression-based functional inferences and that these limitations can be potentially overcome through methodological considerations that include pathway gene annotation quality, expression dataset used, clustering algorithms, and the use of an independent dataset such as the mutant phenotype data used here to maximize the utility of co-expression relationships in hypothesizing gene functional relations.

By evaluating within-pathway gene expression correlation based on the EC measure, we show that genes encoding proteins involved in the same pathway do not necessarily co-express. For example, only 5% of pathways have significantly high EC, using a light treatment dataset. For the remaining 95% of pathways, pathway genes may not be coordinately expressed and/or the light dataset is not informative. For some pathways, co-expression will be ill-suited due to gene sharing among pathways (thus multiple mode of regulation), requirement for condition-specific expression data that is not available, and/or that coordinated regulation of the pathway is at a level beyond transcription. In other situations, several approaches could be taken to improve the recovery of pathways with high EC. By filtering genes based on annotation, it might be possible to obtain a core set of genes that are co-expressed. In addition, using expression datasets of different type (e.g. treatment and/or tissue types), complexity (e.g. individual or combined), and transformation method (e.g. fold change or absolute intensity value) could be effective.

In this study, we have demonstrated that clustering algorithms and parameters impact the ability to find novel pathway genes. Thus, by relying on a single algorithm and a single parameter ─ as is most commonly done in published studies ─ co-expression associations with functional implications might be missed. For any pathway being analyzed it is necessary to find the optimal algorithm and parameters to identify clusters that contain the majority of the known pathway genes. We also demonstrated that using one particular clustering algorithm-parameter combination in most cases does not lead to clusters that have optimal overlaps in gene memberships with pathways. Instead, for the best result, we need to consider multiple algorithms and parameters. The methodological considerations we had in this study reflect the multi-parameter nature of co-expression based analyses. The studies that include co-expression based approaches should involve rigorous testing of multiple variables ranging from the pathway of interest to expression dataset and clustering algorithm.

2.5 Methods

2.5.1 A. thaliana metabolic pathways and pathway features analyzed

A. thaliana metabolic pathways (AraCyc pathways), the genes belonging to these pathways and supporting evidence were obtained from the Plant Metabolic Network (version 8, [63]). To examine a broader set of gene function in addition to metabolism, *A. thaliana* Gene Ontology biological processes (GO-BPs) annotations were obtained from geneontology.org [64]. Only nuclear genes and pathways/processes with >2 genes were included in further analyses. The metabolic pathway genes were divided into two sets based on supporting evidence. The first set contained all pathway genes regardless of the types of evidence supporting the annotations (382 pathways, 5,991 genes). The second set only contained genes with experimental evidence (225 pathways, 934 genes). For the GO data, we examined 1,710 GO-BP terms covering 23,157 genes.

To determine if genes of a pathway tend to have a particular subcellular location, subcellular location information was obtained from the SUBcellular Arabidopsis consensus database (SUBAcon [65]), and a contingency table for each pathway and subcellular location was established to calculate the enrichment *p*-value (Fisher's Exact Test). The resulting *p*-values were corrected for multiple testing [66]. To determine whether similar sets of *cis*-regulatory elements are present among genes in the same pathway, 349 position frequency matrices taken from the Cis-BP database [40] were converted to position weight matrices (PWMs) based on the *A. thaliana* background AT and CG frequencies (0.33 and 0.17, respectively) using the Tools for Analysis of MOtifs (TAMO) package MotifTools [67]. The PWMs were used to determine the location of motif sites in the 1kb region upstream of the transcriptional start sites of *A. thaliana* genes with Motility [68]. To assess the impact of post-transcriptional regulation, we used a dataset with associations between miRNAs and their target genes, downloaded from The Arabidopsis Information Resource-TAIR [69].

2.5.2 Expression dataset and its processing

Six publicly available Affymetrix ATH1 microarray gene expression datasets used in this study include: Biotic stress: GSE5615-5616, Light: GSE5617, Abiotic stress: GSE5620-5628, Development: GSE5629-5634, Hormone: GSE39384, and Diurnal [32,70–72]. In addition to these, ~700 *A. thaliana* microarray datasets were downloaded from NASCArrays database [50]. The datasets were downloaded in either normalized form [50,73] or as unprocessed data from Gene Expression Omnibus (GEO) [74]. For the unprocessed datasets, the CEL files for the AtGenExpress data [32,70,71] were downloaded from TAIR [69] and quantile normalized using the Bioconductor affy package in R [75]. The Bioconductor LIMMA package [76] was used to calculate fold changes by contrasting treatment and control experiments, and the *p*–values of significant fold changes were corrected for multiple testing [66].

2.5.3 Calculation of expression correlation and expression coherence

To generate the null expression correlation distribution, 500,000 gene pairs were randomly selected and their Pearson Correlation Coefficients (PCCs) were calculated using the SciPy library [77]. The 95th percentile PCC values (PCC95) in the null distributions were used as thresholds for calling the expression patterns of two genes as significantly correlated with a 5% false positive rate. Using the PCC95 values, the expression coherence (EC) score was calculated to determine the extent of co-expression among genes in a given pathway [24,25]. The EC score of a pathway is the ratio of the number of gene pairs with PCC values higher than PCC95 and the total number of gene pairs in a pathway. Thus EC values range from 0 (no gene pair with significant expression correlation) to 1 (all gene pairs significantly co-expressed). To identify pathways with significantly higher than randomly expected ECs (high EC pathways), pathway-gene associations were randomized 100 times with the sizes of the pathways kept the same. For each dataset, a distribution of randomly expected EC values was established. For a given dataset, a pathway was defined as a high EC pathway if it had an EC score larger than the 95 percentile value of the null EC distribution. The percentile of the pathway ECs in the null EC distribution was referred to as EC percentile. To assess how similar the gene expression profiles among array experiments in a dataset, the PCC values between the experiments in a dataset were calculated and the median PCC value was used as a measure of homogeneity among the experiments within a dataset.

To evaluate the impact of similarity measures, Spearman's rank coefficient [77], partial correlation [37] and Mutual Information (MI) [78] were used as additional similarity measures to determine pathway EC in the same way as PCC was used. Partial correlations of pathway genes were calculated with two methods: (1) a Python implementation of partialcorr function in MATLAB, which determined the correlations between residuals of linear regression, and (2) the R package corpcor that was optimized for genomic datasets [37]. MI was calculated both as normalized and adjusted with the Python scikit-learn package [78]. The adjusted MI measure accounts for impacts of sample sizes (larger samples might lead to higher MI) and the normalized MI values was calculated by scaling MI values to between 0 and 1. To explore higher order correlations in addition to pairwise ones, Bayesian Networks (BNs) were constructed for each pathway using the bnlearn package in R [79]. Hill-climbing algorithm was used to construct BNs with options for continuous data. The transformed *p*-values (-log(*p*)) of arc strengths between nodes (genes) in BNs were used as measures of gene association strengths that are used similarly as pairwise similarity. Only the transformed *p*-values were used because they were nearly perfectly correlated with arc strengths (r^2=0.9998). BNs were also constructed for randomized pathways to determine threshold *p*-values for each gene association and the thresholds were then applied to determine how many gene pairs in each pathway have above threshold arc strength *p*-values to determine pathway EC.

2.5.4 Co-expression clustering

To determine the impact of the clustering algorithm on the resulting co-expressed gene clusters, we tested *k*-means [80], hierarchical clustering (hclust), *c*-means [81] and Weighted Gene Coexpression Network Analysis (WGCNA) [82] in the R environment and approximate kernel *k*-means [83] in MATLAB. Clustering parameters tested included the numbers of clusters (*k*), distance measures, and hierarchical clustering algorithms for relevant methods. Initially, we attempted to obtain the optimal *k* for clustering the stress expression dataset by obtaining the "elbow plot". After testing 11 *k* values ranging from 5 to 2000 to assess, we realized that the selected *k* was not necessarily the best and the choice of *k* impacts clustering memberships of genes. For distance-based algorithms, three distance measures (Euclidean, radial basis function kernel and 1-PCC) were tested. For hierarchical clustering, we also explored the impact of average, complete and Ward linkage algorithms. For WGCNA, the pickSoftThreshold function was used to determine the ß values based on the scale-free topology model [82]. Commonly used clustering algorithms ─ such as *k*-means ─ are not deterministic, i.e. they may result in a local optimum solution. To evaluate whether multiple runs could result in significantly different results, we ran *k*-means, approximate kernel *k*-means and *c*-means 10 times. We refer to the similarity among 10 runs as consistency between the runs. In contrast, for hierarchical clustering and WGCNA, the co-expression cluster membership was always the same for every run.

2.5.5 Assessing the overlap in pathway and cluster memberships

Fisher's exact test was used to assess how well memberships within a cluster overlap with those in a pathway. The resulting *p*-values were corrected for multiple testing [66]. For each clustering algorithm-parameter combination, an "over-representation score" between a cluster and a pathway was defined as the -log(*q*) value where a higher score indicates a more significant degree of overlap between cluster and pathway memberships. An over-representation score ≥1.3 (q <0.05) was considered to be statistically significant. To account for the possibility that over-representation of some pathways is spurious we asked how often significant over-representation scores arise from randomized expression data. Specifically, the stress expression dataset was permuted to generate 15 random datasets that were used in *k*-means clustering (*k*=5 to 2000, 10 independent runs for each *k* and each random dataset). The same approach outlined above was also used to assess how well memberships in a pathway overlap with those in a random cluster. Among 1,650 random clusters, none had a significant over-representation score with *A. thaliana* pathways.

To further assess if cluster membership can serve to predict pathway membership, we calculated the F measure (the harmonic mean of precision and recall) for each cluster-pathway combination. Precision is the proportion of correct predictions over total predictions; in our case it was the ratio between the number of genes in a cluster that were also found in a pathway and the total number of genes in that cluster. Recall is the proportion of correct predictions over total true positives; in our case it was the ratio between the number of genes in a cluster that were also found in a pathway and the total number of genes in that pathway. F measure was calculated for each pathway-cluster combination with an over-representation score ≥1.3.

2.5.6 Using phenomics data to evaluate co-expression associations

Here we used the mutant profile data from Chloroplast 2010, a database consisting of phenotypic screening results for mutants of more than 5,000 genes [59,60] to confirm the potential functional links between genes found in the same co-expression cluster. This database includes measurements of amino acids and fatty acids as well as the chloroplast morphology and photosynthetic parameters. Taking leucine degradation as an example, we expected the leucine content to be more similar between mutants of leucine degradation genes and mutants of genes found in the same co-expression cluster than to wild-type and mutants of random genes. To determine whether this was the case, we retrieved the leucine measurements (in nmol/g fresh weight) of 12 leucine annotated degradation genes, genes that were clustered with pathway genes with an over-representation score >1.3 (q <0.05), 1000 random genes, where homozygous T-DNA insertions were available, and 184 wild type control plants included in the Chloroplast 2010 database. Significant differences between the leucine levels of mutants and controls, included randomly selected mutants of genes that are not in the leucine degradation pathway and wild-type plants, were identified with Mann-Whitney tests.

2.6 Acknowledgements

We thank Kevin Childs for his help with WGCNA and advice on clustering, Xi Yin for assistance with approximate kernel *k*-means in MATLAB, Nicholas Panchy and Zing Tsung-Yeh Tsai for assistance with Bayesian Network, and members of the Last and Shiu lab for helpful discussions.

APPENDIX

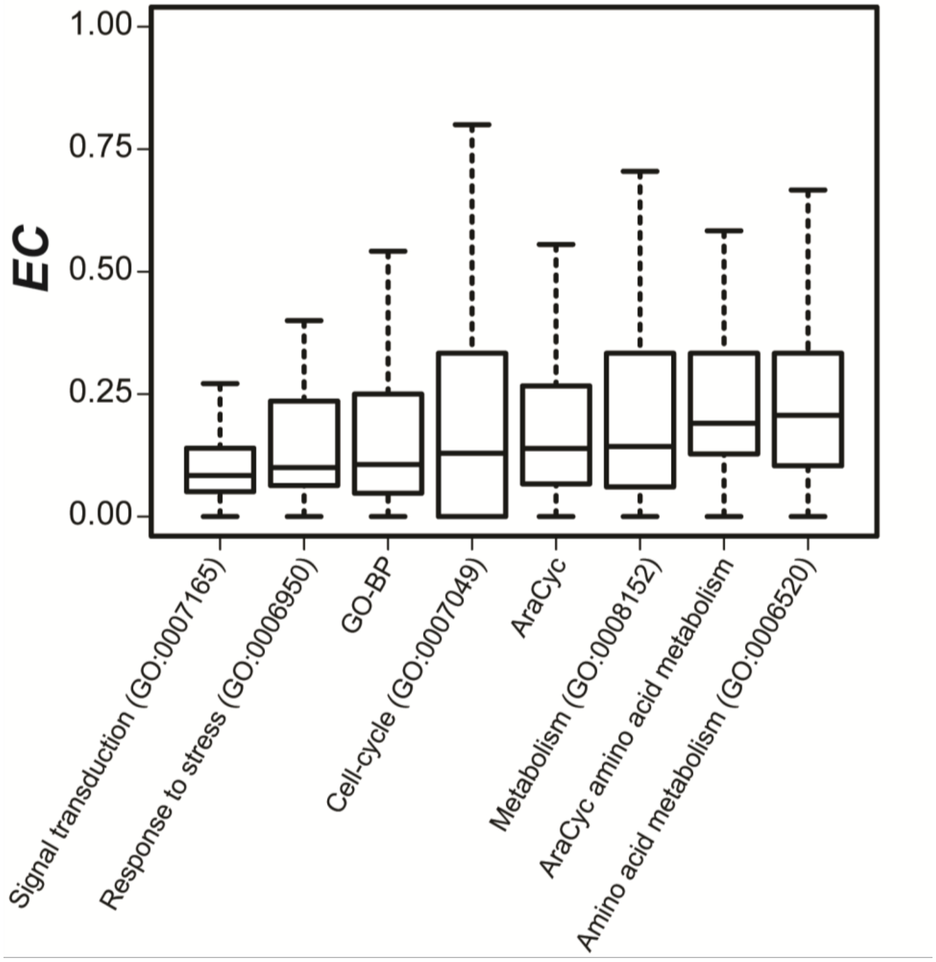


Figure S.2.1 Boxplots of EC values distributions of overall and selected GO-BPs and AraCyc pathways. The order of *x*-axis is based on the median EC.

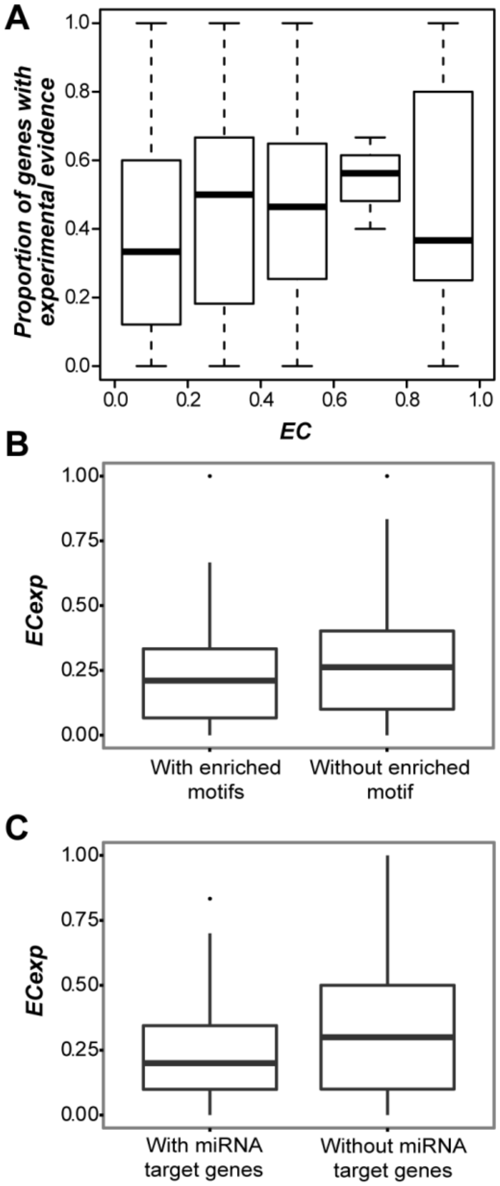


Figure S.2.2 Factors that potentially influence pathway EC. (A) Relationship between the proportion of genes with experimental evidence and EC. EC is shown on the *x*-axis. **(B)** EC calculated for pathway genes that are annotated based on experimental evidence (ECexp) distribution of pathways with and without enriched motifs. **(C)** ECexp distribution of pathways that have miRNA target genes and of pathways those do not.

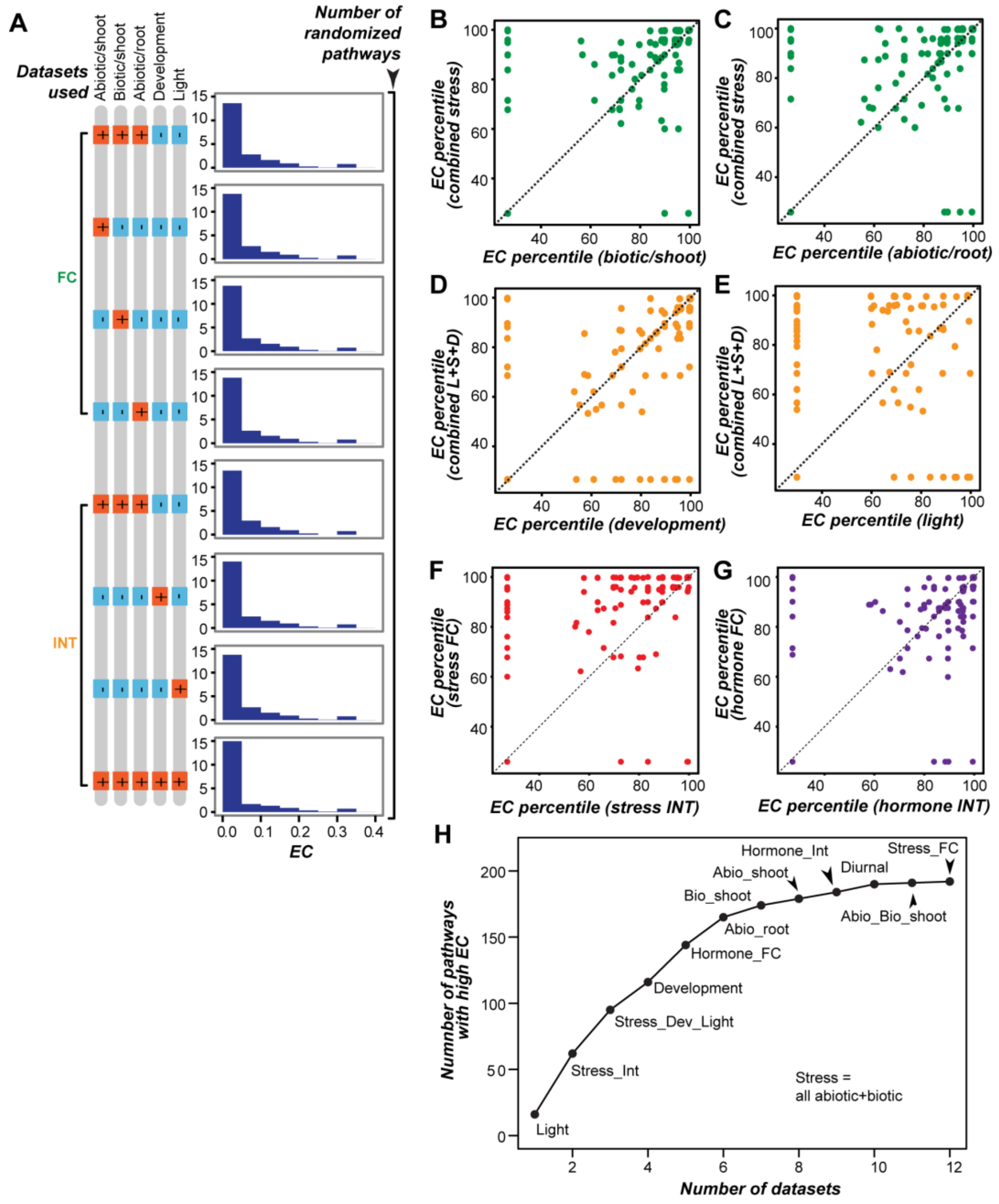


Figure S.2.3 Randomized pathway EC distributions and EC percentiles from multiple datasets. (A) Left panel: Individual and combinations of datasets used to determine pathway EC percentiles. A "+" indicates that the dataset in question was used (either individually or in combination) for the distributions depicted in Right panel.

Figure S.2.3 (cont'd)

Right panel: Distribution of randomized pathway EC. Pathway gene membership was randomized 100 times. Median EC per pathway size is shown in this distribution. **(B)** Relationship between pathway EC percentiles calculated using the combined stress gene expression dataset and those calculated based on individual stress dataset, biotic/shoot. **(C)** abiotic/shoot. **(D)** Relationship between pathway EC percentiles calculated using the light, development, and stress combined dataset and those calculated based on individual dataset, development. **(E)** light. **(F)** Relationship between pathway EC percentiles calculated using the fold change and absolute intensities for the stress gene expression dataset. **(G)** Relationship between pathway EC percentiles calculated using fold change and absolute intensities for the hormone gene expression dataset. Dashed line: *y*=*x*. Each dot represents a pathway. **(H)** The change in the number of pathways with high EC (*y*-axis) with the addition of more expression datasets (*x*-axis).

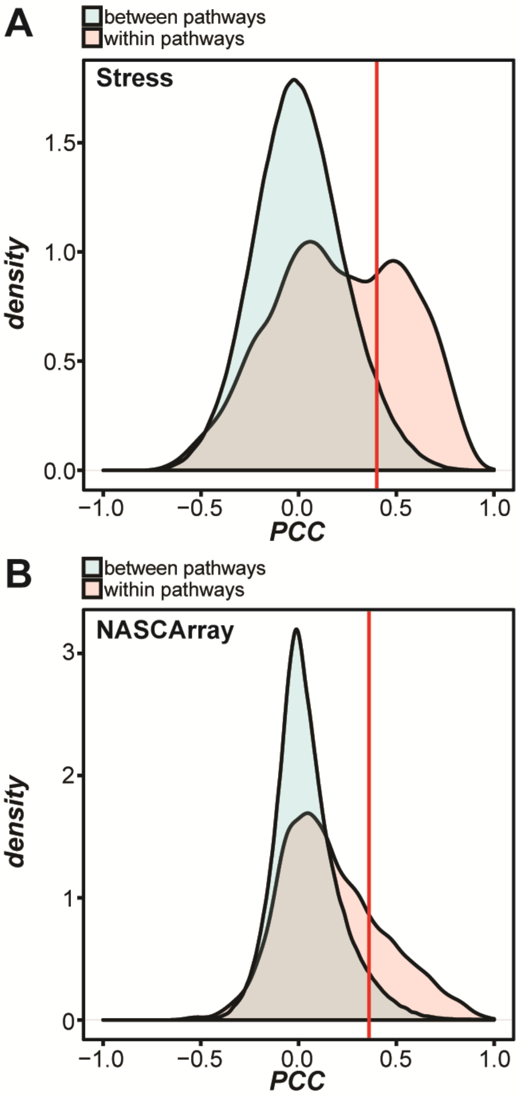


Figure S.2.4 Distinguishing gene pairs within and between pathways using a condition-dependent dataset and a condition-independent dataset. (A) Distributions of PCC values of within pathway gene pairs (light red) and between pathway gene pairs (light blue) using the condition-dependent, stress dataset. Red line: PCC at the 95th percentile of between pathway gene pair PCC distribution. **(B)** Same as (A) using the condition-independent, NASCArray dataset.

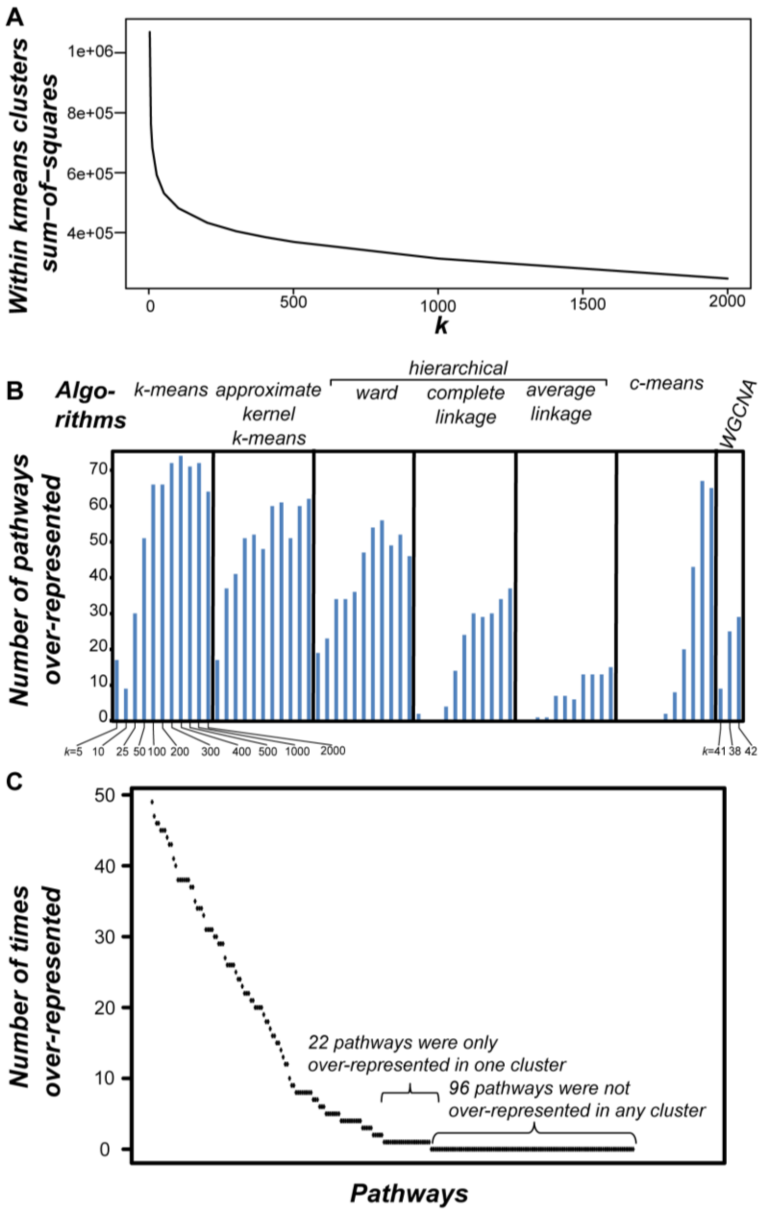
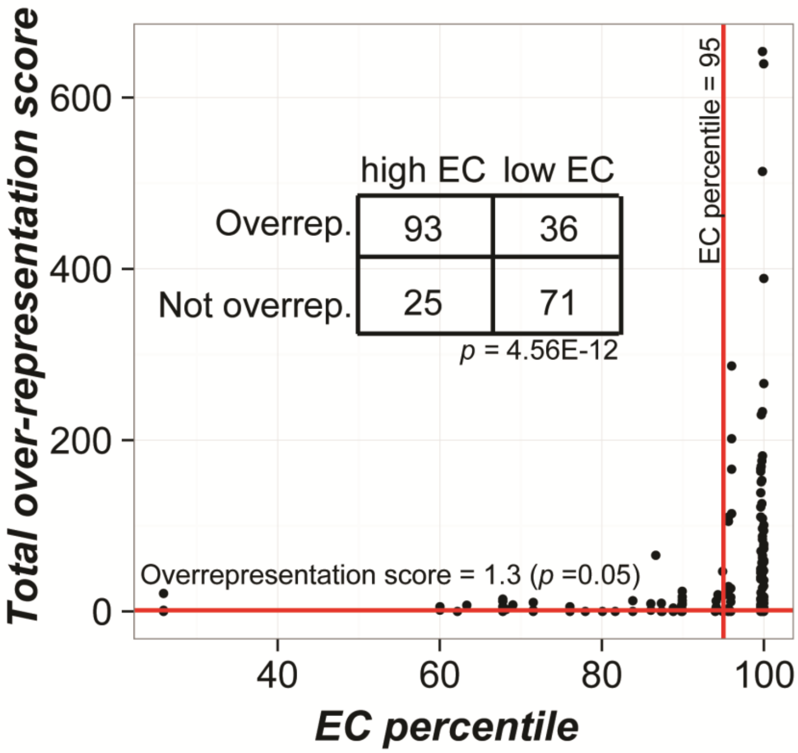


Figure S.2.5 The extent to which metabolic pathways are over-represented in co-expression clusters. (A) Elbow plot showing within cluster sum of squares for *k* = 5-2000. **(B)** The number of pathways over-represented (*y*-axis) in clusters obtained using different algorithms and cluster numbers (*x*-axis). **(C)** Distribution of the number of times that a pathway is over-represented in a cluster (sorted high to low).



	high EC	low EC
Overrep.	93	36
Not overrep.	25	71

Figure S.2.6 The extent to which high EC pathways are over-represented in clusters. The relationship between the pathway-cluster over-representation score (y-axis) and pathway EC percentile (x-axis). Horizontal red line: over-representation score=1.3, corresponding to adjusted *p*-value of 0.05. A pathway is considered over-represented if the over-representation score is >1.3. Vertical red line: EC percentile=95. High EC pathway has an EC percentile>95. The insert is the contingency table for testing (Fisher's exact test) whether the number of high EC pathways is higher among over-represented clusters than randomly expected.

REFERENCES

REFERENCES

1. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40: D1202-10. doi:10.1093/nar/gkr1090

2. Rhee SY, Mutwil M. Towards revealing the functions of all genes in plants. Trends Plant Sci. Elsevier Ltd; 2014;19: 212–221. doi:10.1016/j.tplants.2013.10.006

3. Rosa BA, Jasmer DP, Mitreva M. Genome-wide tissue-specific gene expression, co-expression and regulation of co-expressed genes in adult nematode Ascaris suum. Jex AR, editor. PLoS Negl Trop Dis. Public Library of Science; 2014;8: e2678. doi:10.1371/journal.pntd.0002678

4. Provart NJ, Alonso J, Assmann SM, Bergmann D, Brady SM, Brkljacic J, et al. 50 years of Arabidopsis research: highlights and future directions. New Phytol. 2016;209: 921–44. doi:10.1111/nph.13687

5. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. 1998;95: 14863–14868. doi:10.1073/pnas.95.25.14863

6. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998;9: 3273–97.

7. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res. 2002;12: 37–46. doi:10.1101/gr.205602

8. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res. 2004;14: 1085–94. doi:10.1101/gr.1910904

9. Hirai MY, Sugiyama K, Sawada Y, Tohge T, Obayashi T, Suzuki A, et al. Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc Natl Acad Sci U S A. 2007;104: 6478–83. doi:10.1073/pnas.0611629104

10. Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, et al. Transcriptional coordination of the metabolic network in Arabidopsis. Plant Physiol. 2006;142: 762–74. doi:10.1104/pp.106.080358

11. Mentzen WI, Peng J, Ransom N, Nikolau BJ, Wurtele ES. Articulation of three core metabolic processes in Arabidopsis: fatty acid biosynthesis, leucine catabolism and starch metabolism. BMC Plant Biol. 2008;8: 76. doi:10.1186/1471-2229-8-76

12. Guttikonda SK, Trupti J, Bisht NC, Chen H, An Y-QC, Pandey S, et al. Whole genome co-expression analysis of soybean cytochrome P450 genes identifies nodulation-specific P450 monooxygenases. BMC Plant Biol. 2010;10: 243. doi:10.1186/1471-2229-10-243

13. Childs KL, Davidson RM, Buell CR. Gene coexpression network analysis as a source of functional annotation for rice genes. PLoS One. 2011;6: e22196. doi:10.1371/journal.pone.0022196

14. Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, et al. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell. 2011;23: 895–910. doi:10.1105/tpc.111.083667

15. Han X, Yin L, Xue H. Co-expression analysis identifies CRC and AP1 the regulator of Arabidopsis fatty acid biosynthesis. J Integr Plant Biol. 2012;54: 486–99. doi:10.1111/j.1744-7909.2012.01132.x

16. Wong DCJ, Sweetman C, Ford CM. Annotation of gene function in citrus using gene expression information and co-expression networks. BMC Plant Biol. 2014;14: 186. doi:10.1186/1471-2229-14-186

17. Righetti K, Vu JL, Pelletier S, Vu BL, Glaab E, Lalanne D, et al. Inference of Longevity-Related Genes from a Robust Coexpression Network of Seed Maturation Identifies Regulators Linking Seed Storability to Biotic Defense-Related Pathways. Plant Cell. 2015;27: 2692–708. doi:10.1105/tpc.15.00632

18. Alexeyenko A, Sonnhammer ELL. Global networks of functional coupling in eukaryotes from comprehensive data integration. Genome Res. 2009;19: 1107–16. doi:10.1101/gr.087528.108

19. Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotechnol. Nature Publishing Group; 2010;28: 149–56. doi:10.1038/nbt.1603

20. Kotera M, Yamanishi Y, Moriya Y, Kanehisa M, Goto S. GENIES: gene network inference engine based on supervised analysis. Nucleic Acids Res. 2012;40: W162-7. doi:10.1093/nar/gks459

21. Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, et al. AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. Nucleic Acids Res. 2015;43: D996-1002. doi:10.1093/nar/gku1053

22. Bergmann S, Ihmels J, Barkai N. Similarities and differences in genome-wide expression data of six organisms. PLoS Biol. 2004;2: E9. doi:10.1371/journal.pbio.0020009

23. Lelli KM, Slattery M, Mann RS. Disentangling the many layers of eukaryotic transcriptional regulation. Annu Rev Genet. Annual Reviews; 2012;46: 43–68. doi:10.1146/annurev-

genet-110711-155437

24. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet. 2001;29: 153–9. doi:10.1038/ng724

25. Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y. Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol. 2009;150: 535–46. doi:10.1104/pp.109.136028

26. Rest JS, Wilkins O, Yuan W, Purugganan MD, Gurevitch J. Meta-analysis and meta-regression of transcriptomic responses to water stress in Arabidopsis. Plant J. 2016;85: 548–560. doi:10.1111/tpj.13124

27. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics. 2015;31: 2123–30. doi:10.1093/bioinformatics/btv118

28. Cosgrove EJ, Gardner TS, Kolaczyk ED. On the choice and number of microarrays for transcriptional regulatory network inference. BMC Bioinformatics. 2010;11: 454.

29. Pirooznia M, Yang JY, Yang MQ, Deng Y, Guyon I, Weston J, et al. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. BioMed Central; 2008;9 Suppl 1: S13. doi:10.1186/1471-2164-9-S1-S13

30. Gerstein M, Jansen R. The current excitement in bioinformatics—analysis of whole-genome expression data: how does it relate to protein structure and function? Curr Opin Struct Biol. 2000;10: 574–584. doi:10.1016/S0959-440X(00)00134-2

31. Pirim H, Seker S. Ensemble Clustering for Biological Datasets. 2012; Available: http://www.researchgate.net/publication/236622906_Ensemble_Clustering_for_Biological_Datasets/file/50463518772638378a.pdf

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

33. Mueller LA, Zhang P, Rhee SY. AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol. 2003;132: 453–60. doi:10.1104/pp.102.017236

34. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4: Article17. doi:10.2202/1544-6115.1128

35. Jones CE, Brown AL, Baumann U. Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics. 2007;8: 170. doi:10.1186/1471-2105-8-170

36. Patel R V, Nahal HK, Breit R, Provart NJ. BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. Plant J. 2012;71: 1038–50. doi:10.1111/j.1365-313X.2012.05055.x

37. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol. 2005;4: Article32. doi:10.2202/1544-6115.1175

38. Li S, Ehrhardt DW, Rhee SY. Systematic analysis of Arabidopsis organelles and a protein localization database for facilitating fluorescent tagging of full-length Arabidopsis proteins. Plant Physiol. 2006;141: 527–39. doi:10.1104/pp.106.078881

39. Mao L, Van Hemert JL, Dash S, Dickerson JA. Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics. BioMed Central; 2009;10: 346. doi:10.1186/1471-2105-10-346

40. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. 2014;158: 1431–1443. doi:10.1016/j.cell.2014.08.009

41. Tsai ZT-Y, Shiu S-H, Tsai H-K. Contribution of Sequence Motif, Chromatin State, and DNA Structure Features to Predictive Models of Transcription Factor Binding in Yeast. PLoS Comput Biol. 2015;11: e1004418. doi:10.1371/journal.pcbi.1004418

42. Kunst L, Samuels AL. Biosynthesis and secretion of plant cuticular wax. Prog Lipid Res. 2003;42: 51–80. doi:10.1016/S0163-7827(02)00045-0

43. Go YS, Kim H, Kim HJ, Suh MC. Arabidopsis Cuticular Wax Biosynthesis Is Negatively Regulated by the DEWAX Gene Encoding an AP2/ERF-Type Transcription Factor. Plant Cell. 2014;26: 1666–1680. doi:10.1105/tpc.114.123307

44. Seo PJ, Lee SB, Suh MC, Park M-J, Go YS, Park C-M. The MYB96 transcription factor regulates cuticular wax biosynthesis under drought conditions in Arabidopsis. Plant Cell. 2011;23: 1138–52. doi:10.1105/tpc.111.083485

45. Ljung K. Auxin metabolism and homeostasis during plant development. Development. 2013;140: 943–50. doi:10.1242/dev.086363

46. Pencík A, Simonovik B, Petersson S V, Henyková E, Simon S, Greenham K, et al. Regulation of auxin homeostasis and gradients in Arabidopsis roots through the formation of the indole-3-acetic acid catabolite 2-oxindole-3-acetic acid. Plant Cell. 2013;25: 3858–70. doi:10.1105/tpc.113.114421

47. Kasahara H. Distinct Isoprenoid Origins of cis- and trans-Zeatin Biosyntheses in Arabidopsis. J Biol Chem. 2004;279: 14049–14054. doi:10.1074/jbc.M314195200

48. Feng H, An F, Zhang S, Ji Z, Ling H-Q, Zuo J. Light-regulated, tissue-specific, and cell differentiation-specific expression of the Arabidopsis Fe(III)-chelate reductase gene AtFRO6. Plant Physiol. 2006;140: 1345–54. doi:10.1104/pp.105.074138

49. Kim SA, Guerinot M Lou. Mining iron: iron uptake and transport in plants. FEBS Lett. 2007;581: 2273–80. doi:10.1016/j.febslet.2007.04.043

50. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res. Oxford University Press; 2004;32: D575-7. doi:10.1093/nar/gkh133

51. Bilgin DD, Zavala JA, Zhu J, Clough SJ, Ort DR, DeLucia EH. Biotic stress globally downregulates photosynthesis genes. Plant Cell Environ. 2010;33: 1597–613. doi:10.1111/j.1365-3040.2010.02167.x

52. D'haeseleer P. How does gene expression clustering work? Nat Biotechnol. Nature Publishing Group; 2005;23: 1499–501. doi:10.1038/nbt1205-1499

53. Pelleg DD, Pelleg DD, Moore AW, others. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. ICML. 2000. pp. 727--734. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.3377

54. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B (Statistical Methodol. Blackwell Publishers Ltd.; 2001;63: 411–423. doi:10.1111/1467-9868.00293

55. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. North-Holland; 1987;20: 53–65. doi:10.1016/0377-0427(87)90125-7

56. Ben-Hur A, Elisseeff A, Guyon I. A stability based method for discovering structure in clustered data. Pacific Symp Biocomput. 2002;7: 6–17.

57. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010;31: 651–666. doi:10.1016/j.patrec.2009.09.011

58. Pollard KS, van der Laan MJ. Cluster Analysis of Genomic Data. Springer New York; 2005. pp. 209–228. doi:10.1007/0-387-29362-0_13

59. Lu Y, Savage LJ, Larson MD, Wilkerson CG, Last RL. Chloroplast 2010: a database for large-scale phenotypic screening of Arabidopsis mutants. Plant Physiol. 2011;155: 1589–600. doi:10.1104/pp.110.170118

60. Bell SM, Burgoon LD, Last RL. MIPHENO: data normalization for high throughput metabolite analysis. BMC Bioinformatics. 2012;13: 10. doi:10.1186/1471-2105-13-10

61. Peng C, Uygun S, Shiu S-H, Last RL. The Impact of the Branched-Chain Ketoacid Dehydrogenase Complex on Amino Acid Homeostasis in Arabidopsis. Plant Physiol. 2015; pp.15.00461-. doi:10.1104/pp.15.00461

62. Gu L, Jones AD, Last RL. Broad connections in the Arabidopsis seed metabolic network revealed by metabolite profiling of an amino acid catabolism mutant. Plant J. 2010;61: 579–90. doi:10.1111/j.1365-313X.2009.04083.x

63. Plant metabolic pathway database (PMN / PlantCyc) home page [Internet]. [cited 12 Apr 2016]. Available: http://www.plantcyc.org/

64. Gene Ontology Consortium TGO. Gene Ontology Consortium: going forward. Nucleic Acids Res. Oxford University Press; 2015;43: D1049-56. doi:10.1093/nar/gku1179

65. Hooper CM, Tanz SK, Castleden IR, Vacher MA, Small ID, Millar AH. SUBAcon: a consensus algorithm for unifying the subcellular localization data of the Arabidopsis proteome. Bioinformatics. 2014; btu550-. doi:10.1093/bioinformatics/btu550

66. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100: 9440–5. doi:10.1073/pnas.1530509100

67. Gordon DB, Nekludova L, McCallum S, Fraenkel E. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics. 2005;21: 3164–5. doi:10.1093/bioinformatics/bti481

68. Cartwheel: a framework for genomic sequence analysis [Internet]. [cited 12 Apr 2016]. Available: http://cartwheel.caltech.edu/

69. TAIR. [cited 12 Apr 2016]. Available: http://www.arabidopsis.org/

70. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, et al. A gene expression map of Arabidopsis thaliana development. Nat Genet. 2005;37: 501–6. doi:10.1038/ng1543

71. Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, et al. The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access. Plant J. 2008;55: 526–42. doi:10.1111/j.0960-7412.2008.03510.x

72. Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, et al. The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. Cold Spring Harb Symp Quant Biol. 2007;72: 353–63. doi:10.1101/sqb.2007.72.006

73. AtGenExpress Resources - Weigel World [Internet]. [cited 12 Apr 2016]. Available: http://jsp.weigelworld.org/AtGenExpress/resources/

74. GEO - NCBI. [cited 12 Apr 2016]. Available: http://www.ncbi.nlm.nih.gov/geo/

75. Bioconductor. [cited 12 Apr 2016]. Available: http://www.bioconductor.org/

76. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3: Article3. doi:10.2202/1544-6115.1027

77. SciPy.org. [cited 12 Apr 2016]. Available: http://www.scipy.org/

78. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

79. bnlearn-R package. Available: http://www.bnlearn.com/

80. Hartigan J, Wong M. Algorithm AS 136: A K-Means Clustering Algorithm. Appl Stat. 1979;28: 100. doi:10.2307/2346830

81. Pal NR, Bezdek JC, Hathaway RJ. Sequential Competitive Learning and the Fuzzy c-Means Clustering Algorithms. Neural Networks. Elsevier Science Ltd.; 1996;9: 787–796. doi:10.1016/0893-6080(95)00094-1

82. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9: 559. doi:10.1186/1471-2105-9-559

83. Chitta R, Jin R, Havens TC, Jain AK. Approximate Kernel *k*-means : Solution to Large Scale Kernel Clustering. 2011; 895–903.

CHAPTER 3

PREDICTIVE MODELS OF SPATIAL TRANSCRIPTIONAL RESPONSE TO HIGH SALINITY[1]

[1]The work described in this chapter was submitted to the journal *Plant Physiology* and currently under review for publication:

Sahra Uygun[¶], Alexander E. Seddon[¶], Christina B. Azodi, Shin-Han Shiu. Predictive models of spatial transcriptional response to high salinity

[¶] These authors contributed equally to this work.

Contribution: Alexander E. Seddon and I were involved in conceptualization, data curation and formal analyses of this project. We generated all figures of this manuscript. Alexander E. Seddon and I wrote the original manuscript, I updated and included sections, and all authors were involved in reviewing and editing of the manuscript.

3.1 Abstract

Plants are exposed to a variety of environmental conditions, and their ability to respond to environment variation depends on the proper regulation of gene expression in an organ, tissue, and cell type specific manner. Although knowledge is accumulating on how stress responses are regulated, a genome-wide model of how plant transcription factors (TFs) and *cis*-regulatory elements (CREs) control spatially specific stress response has yet to emerge. Using *Arabidopsis thaliana* as a model, we identified a set of 1,894 putative CREs (pCREs) that are associated with high salinity (salt) up-regulated genes in the root or the shoot. These pCREs led to computational models that can better predict salt up-regulated genes in root and shoot compared to models based on known TF binding motifs. In addition, we incorporated TF binding sites identified via large-scale *in vitro* assays, chromatin accessibility, evolutionary conservation and pCRE combinatorial relations in machine learning models, and found that only consideration of pCRE combinations led to better performance in salt up-regulation predictions in root and shoot. Our results suggest that the plant organ transcriptional response to high salinity is regulated by a core set of pCREs and provide a genome-wide view on the *cis*-regulatory code of plant spatial transcriptional responses to environmental stress.

3.2 Introduction

Plants are equipped with a wide range of mechanisms to respond to environmental stresses such as excess heat, salinity, drought, and pathogen attack [1,2]. These stress response mechanisms are indispensable for plant survival and have a significant spatial component where organs and tissues respond to the environmental changes differently [3–5]. In the case of high salinity stress (referred to as salt stress), after perceiving an increase in soil salt concentration, the primary physiological response of the root is to exclude sodium from the xylem and to send hormonal signals of stress to the shoot, while the shoot must respond to the effects of ion toxicity and water limitation [6,7]. In addition to physiological changes that are spatially specific, it is well documented that differential gene expression under stress conditions can be regulated in highly organ and tissue specific manner [8–11], which ultimately impact plant development and physiology.

Spatially- and conditionally- specific gene expression is expected to be subjected to the control of transcriptional regulatory machineries, including transcription factors (TFs) and their associated *cis*-regulatory elements (CREs). Currently, the TFs and their corresponding CREs regulating stress response have received considerable attention [12–14], but our knowledge on spatial regulation of stress responses are limited. CREs can be identified based on co-expression [15–19] and/or through *in vitro* and *in vivo* TF binding experiments [20–23]. The co-expression approach has been successfully used to identify putative CREs (pCREs) in regulating stress responsive gene expression in yeast (*Saccharomyces cerevisiae*) [15] and in *A. thaliana* [17]. In addition, pCREs are over-represented in the 1kb regions upstream of tissue and cell type specifically expressed genes [24]. Although some of these pCREs are similar to binding sequences of TFs known to regulate stress responsive genes [24], it remains unclear how they may be relevant to spatial stress response regulation. One fruitful computational approach for assessing the relevance of pCREs is to ask how well they can be used to predict spatial stress response, i.e. how well they can be used to identify the "*Cis*-regulatory Code" (CRC [17,25])

CRC is defined as the sets of CREs involved in gene regulation in a particular context (e.g. environment, location, timing) [17,25]. One major conclusion from CRC studies is that TFs frequently regulate genes expression pattern in combinations. For example, in yeast, CREs identified through TF binding data uncovered a complex regulatory code involving combinations of multiple CREs [20]. In humans, genes expressed in specific tissues are regulated by particular combinations of TFs and CREs [26]. In *A. thaliana*, CRCs consisting of binary combinations of pCREs resulted in more precise predictions of salt stress up-regulated genes [17] then using individual CREs. CRCs can potentially be further improved by knowledge of TF binding. For example, computational model considering *in vitro* TF binding site information, sequence conservation, DNA structure, and/or chromatin accessibility were shown to predictive of *in vivo* TF binding in mouse [27] and in yeast [28]. Tissue specific TF binding was also predicted using information on binding motifs and histone modifications [29]. These examples highlight the relevance and utility of CRCs and integration of multiple relevant datasets for understanding the mechanisms underlying genome-wide spatial transcriptional response to stress. However, such spatial response CRC is not available.

The goal of this study was to uncover the CRCs underlying spatially specific transcriptional response to stress using plant as a model. Specifically, we focused on the CREs relevant to salt stress response in the above ground (shoot) and the below ground (root) parts of *A. thaliana*. Salt stress was chosen because it is well studied both physiologically [6,7] and molecularly [30,31], and there are documented differences in the transcriptional response to salt in the root and shoot [8,11]. Additionally, there are known TFs and CREs for salt stress [31–34] for verifying our results. To assess transcriptional changes to salt stress across different organs in *A. thaliana*, we first asked how salt up-regulated genes in roots and shoots differed in their functional annotations. Next, to determine how well current knowledge of TF binding sites in *A. thaliana* can explain spatial salt up-regulation, we used motifs and binding sites identified through two large-scale *in vitro* studies [22,23] to generate models of root and shoot salt up-regulation. We then identified additional putative CREs (pCREs) with a co-expression approach to assess if these newly identified pCREs allowed better predictions of spatial response to salt stress. We tested whether pCRE individually could be used to establish a *cis*-regulatory model explaining spatial patterns of up-regulation during salt stress. To evaluate whether we could further improve spatial salt stress response prediction, we filtered pCRE sites according to information on *in vitro* TF binding [22,23], chromatin accessibility, and conserved non-coding regions [35]. Lastly, we built prediction models using combinations of CREs.

3.3 Results and Discussion

3.3.1 Transcriptional responses to stress have a strong spatial component

Earlier global gene expression studies demonstrated that different plant organs have distinct transcriptional responses to stress [8–11]. To assess the extent to which organs have unique expression patterns under different stress conditions and to determine the similarities between organ (root vs. shoot) stress responses, we determined the correlations between the levels of differential expression across multiple conditions and time points using two types of existing datasets: (1) root and shoot samples under abiotic stress [11] and (2) shoot samples under biotic stress (see **3.5 Methods**). There were several patterns worth noting. First, samples for related stress conditions tended to cluster together, and these "stress condition clusters" tended to have root and shoot sub-clusters (**Figure 3.1A**). For example, osmotic and salt stress samples formed a cluster, with sub-clusters composed of shoot and root samples (dotted rectangles I and II respectively; **Figure 3.1A**).

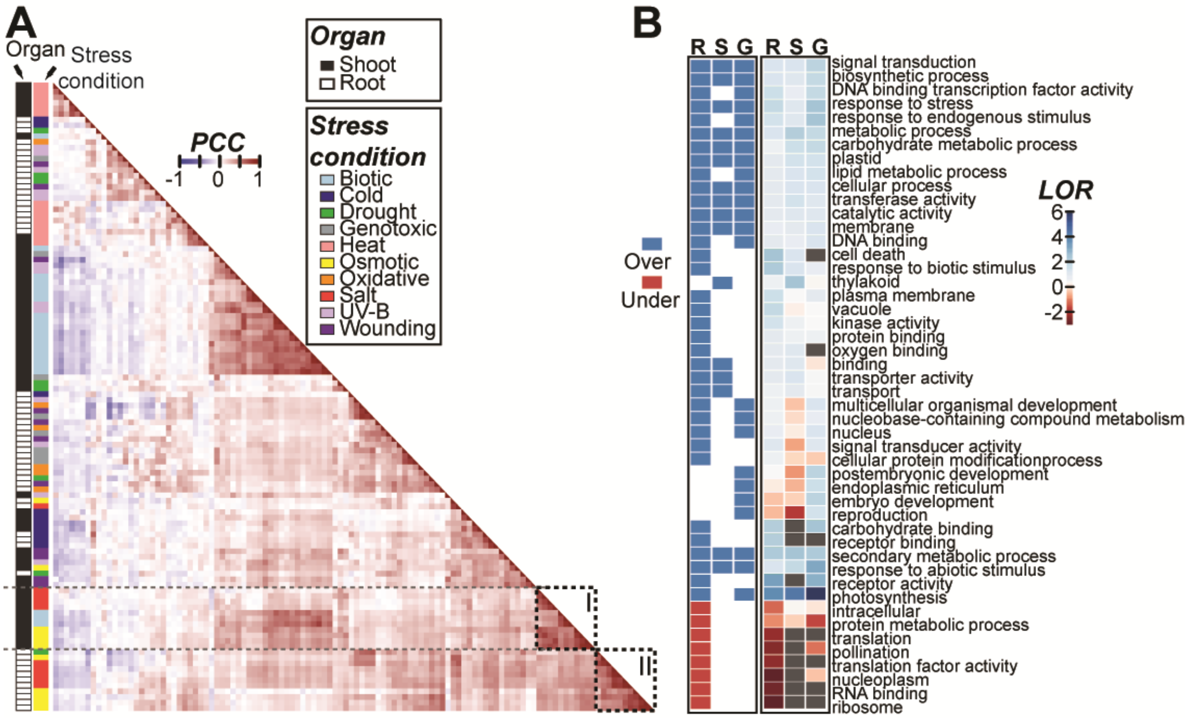


Figure 3.1 *A. thaliana* gene expression correlation across stress datasets and Gene Ontology (GO) terms enriched in salt responsive genes. (A) Between-sample Pearson's Correlation Coefficients (PCC) calculated based on $\log_2$ fold changes ($\log_2$(stress treatment/control)) of genes in shoot and root samples under each stress condition/treatment duration combination. The orders of rows and columns are the same, and they are sorted based on hierarchical clustering of the pairwise PCC values. Boxes on the left represent the key for organ and stress condition. Dotted rectangles I and II highlight osmotic and salt stress clusters, respectively. **(B)** Heatmap (left) indicating Gene Ontology (GO) Slim terms significantly over- (blue) or under- (red) represented in genes that are differentially up-regulated during salt stress after 3 hours in (R)oot only, (S)hoot only, and/or (G)lobally in both organs ($\log_2$ fold change>1, $p \leq 0.05$). Right heatmap summarizes the $\log_2$ odds ratio (LOR) from the enrichment test (grey: LOR could not be calculated due to 0 in ratio).

The median Pearson's Correlation Coefficient (PCC) of the $\log_2$ fold-change values between samples from the same organ but different stress conditions (median PCC=0.17) were significantly lower than those between samples from the same stress condition but different organs (median PCC=0.31, Mann-Whitney, $p<2.2e-16$). Thus, the stress condition has more impact on overall expression pattern than organ identity. Nonetheless, under some stress conditions there were stronger organ specific effects. For example, the salt stress response correlations between organs (median PCC=0.24) were significantly lower those between samples from the same organ (median PCC=0.69, Mann-Whitney, $p<2.2e-16$). Taken together, our findings are consistent with earlier studies [8–10] that, while there is a specific transcriptional response to each stress, this response is further influenced by the organ where genes are expressed.

Given the stress response is influenced by organ identity, we next assessed what types of genes tend to be differentially up-regulated in root and shoot during salt stress using Gene Ontology (GO) term enrichment analysis (see **3.5 Methods**). Three sets of significantly salt up-regulated genes were defined: (1) global - 246 genes both in the root and the shoot; (2) root-specific - 1,854 genes only in the root, and (3) shoot-specific - 276 genes only in the shoot. There were 48 GO terms significantly over/under-represented in ≥1 gene sets defined above (**Figure 3.1B**). For example, thylakoid and plastid terms were over-represented among shoot specifically up-regulated genes, consistent with an earlier finding that photosynthesis is significantly impacted by salt stress [36]. Among the terms, signal transduction and response to stress were over-represented in all three gene sets (**Figure 3.1B**). As these three gene sets are mutually exclusive, this result suggests that the root and the shoot have unique signaling pathway genes up-regulated, as well as pathways that are globally necessary for stress response. This result is supported by work on the The Salt Overly Sensitive (SOS) pathway that involves components that are common to both organs as well as those specific to the root and shoot [30,37]. Interestingly, “DNA binding transcription factor activity” and "DNA binding", were only enriched amongst the root specifically and globally up-regulated genes. This suggests that there is a set of global TFs and another set specific to the root. In addition, genes up-regulated in the root may be regulated by both a global and a root specific set of TFs, while genes up-regulated in the shoot may be regulated primarily by a global TF set.

To summarize, a variety of functional categories were found to be enriched in the genes up-regulated by salt stress. In some instances, root specifically, shoot specifically, and globally up-regulated genes had the same enriched functional categories. These common enriched terms suggest that roots and shoots up-regulate similar types of genes. However, there are also genes up-regulated in an organ specific manner that may be regulated by distinct sets of up-regulated TFs. The TFs that are specifically up-regulated in roots may help to explain the differences in expression pattern that we observe between the roots and shoots under salt stress. Thus, the root specifically up-regulated genes may be controlled by the root specific TFs. Because TFs may differ in the CREs they bind and there are substantial amounts of *in vitro* TF-DNA interaction data in *A. thaliana* [22,23], we next examined whether known TF binding data may be associated with organ specific, salt induced gene expression.

3.3.2 Known TF binding motifs contribute to a better than random performing model for salt up-regulation prediction

Because TFs exert their regulatory roles by binding to CREs, we expected that the global and organ specific activities of TFs will be reflected in which CREs are in the regulatory regions of global, root-specific, and shoot-specific salt up-regulated genes. We hypothesized that each organ had a different set of CREs regulating its salt stress up-regulated genes and these CREs could be used to construct CRCs that are models for predicting stress responsive gene expression

[17]. To test these hypotheses, we collected *A. thaliana* TF binding data from two large-scale *in vitro* studies, CIS-BP [22] and DAP-seq [23] that cover binding sites of 758 TFs. Given the extensive coverage of TFs, we expected that these datasets would cover a significant number of *cis*-regulatory sequences relevant for controlling root and/or shoot up-regulated genes. Here the root up-regulated genes were defined as the union of the root specifically and globally (in both root and shoot) up-regulated genes under high salinity treatment. Similarly, the shoot up-regulated genes were the union of the shoot specifically and globally up-regulated genes.

We first tested if the TF binding sites predicted based on the CIS-BP data and the binding sites inferred from DAP-seq peaks were significantly over-represented in the putative promoter regions (within 1kb upstream of transcriptional start sites) of root and shoot up-regulated genes. Among binding site information for 758 TFs, we found that 262 and 397 were significantly over-represented in the putative promoters of root and shoot up-regulated genes, respectively compared to non-responsive genes. Overall, we found that, if the TF binding sites based on the CIS-BP data are enriched in the promoter regions of root up-regulated genes, the same sites also tend to be enriched among shoot up-regulated genes (enrichment score PCC=0.88, *p*=9.20e-42; **Figure 3.2A**). This was also the case when DAP-seq data was used, but to a much lesser degree (PCC=0.25, *p*=2.27e-04; **Figure 3.2B**). This finding suggests that some *cis*-regulatory sites are common between root and shoot up-regulated genes. Nonetheless, the correlations were not perfect, suggesting that some CIS-BP TF and DAP-seq binding sites were differentially enriched between up-regulated genes in root and shoot. Consistent with this notion, the binding sites of some TF families were enriched in an organ specific manner. For example, WRKY binding sites were over-represented only in root and AP2 sites were over-represented mostly in shoot up-regulated genes (**Figure 3.2A and B**). Next, to assess the extent to which known TF binding data could explain organ-specific responses, we established CRCs with machine learning methods for predicting whether a gene is up-regulated or non-responsive to salt stress in root or shoot based on the presence and absence of CIS-BP TFBM or DAP-seq sites in the putative promoter regions (see **3.5 Methods**).

We used two approaches to evaluate CRC model performance. The first is Area Under Curve - Receiver Operating Characteristic (AUC-ROC), where a perfect model would have AUC-ROC=1 and random predictions would lead to AUC-ROC=0.5. The second way is the precision-recall curve, where precision is the proportion of correctly predicted genes that are up-regulated in an organ and recall is the proportion of truly up-regulated genes in an organ that are correctly predicted. Better models would have precision-recall curves tending more towards the upper-right corner of the graph and random predictions would be no better than the background (dotted lines, **Figure 3.2C and D**). The model built with all binding site data according to CIS-BP or DAP-seq led to better predictions than randomly expected in root and in shoot (**Figure 3.2C and D**), indicating that these TF binding data contain relevant regulatory information for root and shoot salt up-regulation. Consistent with the expectation that only a subset of TFs would be involved in the organ-specific up-regulation, models using binding data of TFs with over-represented numbers of binding sites in salt up-regulated genes resulted in similar performance as the ones using all TF data in either root or shoot (**Figure 3.2C and D**). In addition, the models based on binding sites of TFs that were not over-represented performed poorly (AUC-ROC=0.54-0.56).

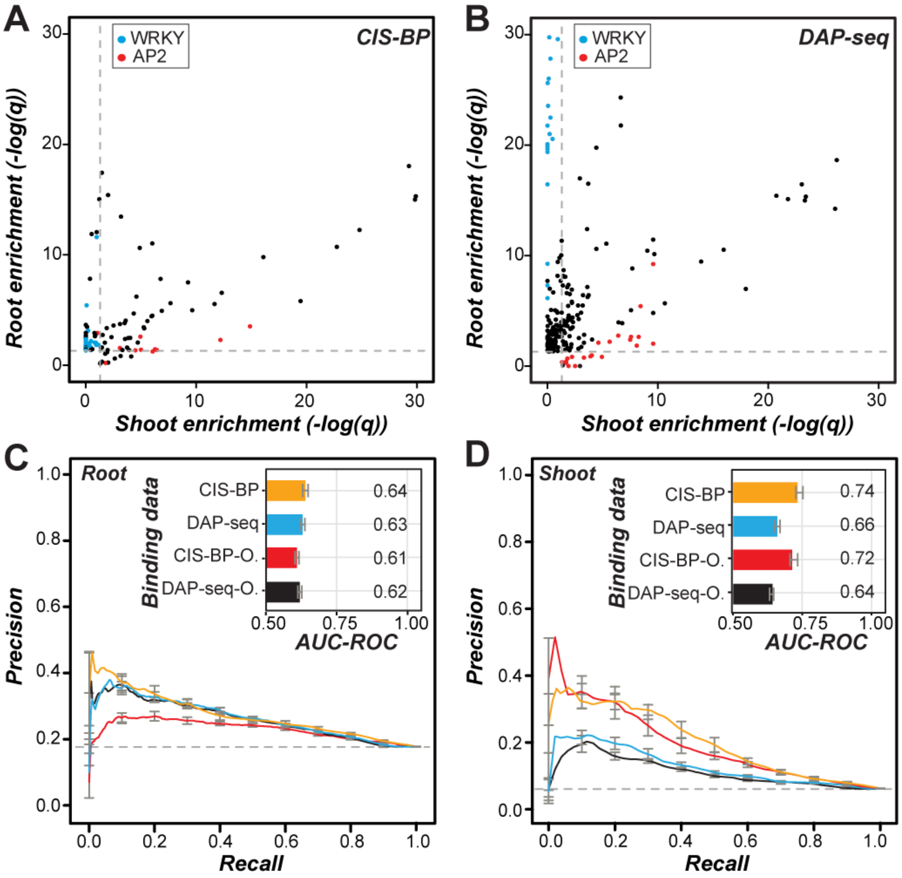


Figure 3.2 Over-representation of known TF binding sites in organ salt up-regulated genes and performance of CRCs predicting salt up-regulation. (A) Scatterplot of enrichment score (–log (*q*-value)) of CIS-BP TFBM sites in the promoters of root (*y*-axis) and shoot (*x*-axis) up-regulated genes compared to non-responsive genes. Each point is for one TFBM. Blue: WRKY TF data. Red: AP2 TF data. Dotted lines: *q*-value threshold at 0.05. **(B)** As in (A) but using DAP-seq data. Each point is for one TF. **(C)** Precision-recall curves and AUC-ROCs (insert) of CRCs predicting root up-regulated genes using CIS-BP TFs (orange) or DAP-seq TFs (blue). O: over-represented among root up-regulated genes (red and black). The colors of the precision-recall curves correspond to the colors for different subsets in the AUC-ROC bar chart. The error bar corresponds to the standard error from 10-fold cross validation for each model. **(D)** As in (C) but for shoot up-regulated genes.

As the TF binding information was available for ~38% of the known *A. thaliana* TFs tested under *in vitro* conditions [22,23,38], there could be some relevant CREs not be included in the models. In addition, it is worth noting that performance of modeling root up-regulated genes is not as good as modeling salt up-regulated genes in shoots (**Figure 3.2C and D**). Thus, to improve upon our understanding of what CREs are associated with and how these CREs may influence salt up-regulation in the root and shoot, we next identified putative CREs based on co-expression to assess how the regulatory logic differs between the root and shoot salt up-regulation.

3.3.3 pCREs derived from co-expression clusters are similar, but not identical, to the known binding motifs of TFs

We hypothesized that motifs identified through co-expression clustering would provide additional regulatory information compared to large-scale TF *in vitro* binding data [22,23] in modeling salt up-regulation. To test this, we identified 1,894 putative CREs (pCREs) over-represented in putative promoters of root and/or shoot salt up-regulated genes in co-expression clusters defined based on the stress fold-change data (see **3.5 Methods**). Next, we asked if the pCREs identified based on co-expression were similar to 355 CIS-BP TF binding motifs (TFBMs) [22]. We calculated the PCC values of the position weight matrices (PWMs) of all motif pairs between pCRE and CIS-BP TFBMs to find the best matching pCRE-TFBM pairs where lower PCC values indicating diminishing similarity (**Figure 3.3A**). Three criteria were used to define whether a pCRE-TFBM pair had significant similarity. First, we identified pCREs that are identical to TFBMs. Only two pCREs were identical (PCC=1) to experimentally determined binding motifs for ATBZIP63 involved in abscisic acid (ABA) biosynthesis [34] and ABF3 involved in abscisic acid (ABA) signaling [39], consistent with their roles in salt stress response. Second, given PCC=1 is highly stringent, we defined that a pCRE-TFBM pair has significant similarity if their PCC is significantly higher than (at the 5% level) PCCs of TF pairs from the same family (red; **Figure 3.3A**). Based on this criterion, 4% of the pCREs were significantly similar to TFBMs. Third, we defined a pCRE-TFBM pair has significant similarity if their PCC is significantly higher than (at the 5% level) PCCs of TF pairs from different families (blue; **Figure 3.3A)**. This is reasonable because the within family TFBM PCC values tend to be higher than between families (**Figure S.3.1**). This criterion allowed us to assess TFs from which families may bind the pCREs. Based on this criterion, 25% pCREs can be assigned to 24 of the 27 TF families and example TFBMs and their best matching pCREs are shown in **Figure 3.3B**.

While 25% of the organ pCREs enriched among salt up-regulated genes are significantly similar to ≥1 TFBMs, what should be made of the remaining 75% of pCREs? One possibility is that these pCREs are TFBMs likely bound to one of the families, just that in the existing TF binding data a close representative is not available. To test this, we asked if the pCREs are more similar to a known TFBM than to sequences randomly drawn from the genome (black dots; **Figure 3.3A**) and found that PCC values between pCREs and their best matching TFBMs are all higher than 95th percentile value in the pCRE-random sequence PCC distribution (**Figure 3.3A**). Thus, all pCREs are more significantly similar to known TFBMs than random sequences. These findings suggest that the pCREs are not simply random sequences pulled from the genome and that the co-expression-based analysis contributed to an expanded set of CREs that are relevant for organ salt up-regulation.

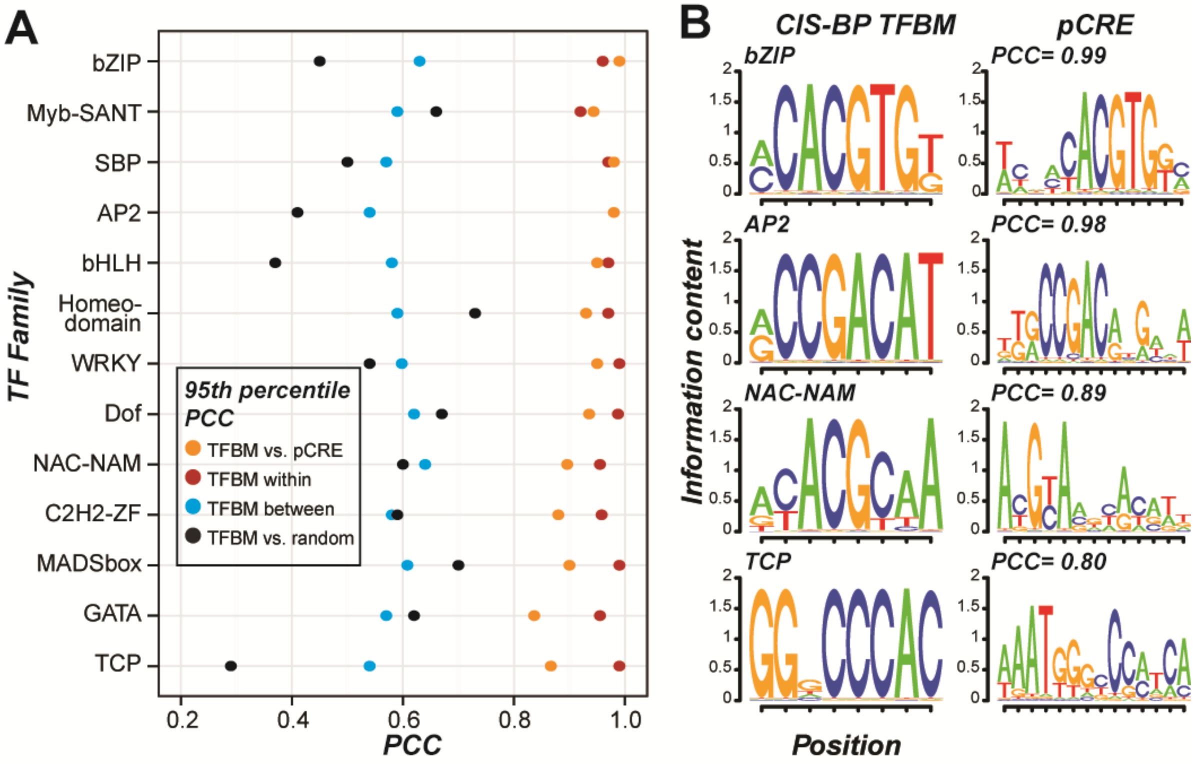


Figure 3.3 Similarity of the pCREs to CIS-BP TFBMs. (A) Distributions of PCC values between TFBMs. The *y*-axis indicates different TF families and the *x*-axis indicates the PCC values. Orange: the 95th percentile value of the distributions of PCCs between TFBMs from a particular family and pCREs with their best matches in the same family (TFBM vs. pCRE) TFBMs. Red: the 95th percentile value in the distribution of PCCs between pairs of TFBMs from each TF family (TFBM within). Blue: the 95th percentile value of the distribution of PCCs between TFBMs in one family and their best matching TFBMs in other families (TFBM between). Black: the 95th percentile value of the distribution of PCCs between TFBM from a particular family and random motifs (TFBM vs. random). **(B)** The sequence logos of bZIP, AP2, NAC-NAM and TCP TFBMs from CIS-BP (left) and their best matching pCREs and PCC values (right).

3.3.4 pCRE set further improves salt up-regulation prediction in a spatially specific manner

To assess if the pCRE set predicts salt up-regulation better than known *in vitro* TF binding sites [22,23], we modeled salt up-regulated expression using the pCRE set (see **3.5 Methods**; **Figure 3.4A and B**). Salt up-regulation prediction models based on pCREs had better prediction performance for both root up-regulated genes (red, AUC-ROC=0.71; **Figure 3.4A**) and shoot up-regulated genes (red, AUC-ROC=0.79; **Figure 3.4B**) than the models based on CIS-BP and DAP-seq data (root AUC-ROC=0.64, shoot AUC-ROC=0.74; **Figure 3.2C and D**). This improvement indicates that using motifs discovered from co-expression clusters containing root and/or shoot up-regulated genes led to better prediction models of organ salt up-regulation.

Next, we classified 1,894 organ pCREs into three subsets that were over-represented in the promoters of genes up-regulated by salt in the root (759 root pCREs), in the shoot (237 shoot pCREs), and in both root and shoot (898 general pCREs). The rationale for defining these pCRE subsets was that the root and shoot subsets might be more critical to controlling expression for the root specifically and shoot specifically up-regulated genes, respectively, while the general pCREs might be critical for globally up-regulated genes. To test this hypothesis, salt up-regulation prediction models were established using root, shoot, and general pCREs where each pCRE was treated as an independent predictor. For predicting root up-regulated genes (including genes up-regulated globally and in root specifically), we found that a model based on root pCREs (AUC-ROC=0.70) was much better than a model based on shoot pCREs (AUC-ROC=0.61; **Figure 3.4A**). Similarly, a model based on shoot pCREs better predicted shoot salt up-regulated genes (AUC-ROC=0.73) than a model based on root pCREs (AUC-ROC=0.66; **Figure 3.4B**). Thus, the root and shoot pCRE sets are better at predicting up-regulated genes in the organs for which they are associated, demonstrating they are relevant to spatially specific up-regulated genes. In addition, root pCREs alone or the combination of the general and the root pCRE set resulted in models that performed as well as the model using the all pCRE set (AUC-ROC=0.71; **Figure 3.4A**). This suggests that shoot pCREs provide no additional information to predict root up-regulated genes. In contrast, for shoot up-regulated genes, although the model based only on shoot pCRE performed reasonably well (AUC-ROC=0.73), it was not as good as the model based only on the general pCREs (AUC-ROC=0.80; **Figure 3.4B**).

Surprisingly, adding the shoot pCRE set did not provide additional regulatory information for salt up-regulation in the shoots that was not already provided by the general pCREs. This conclusion is based on the finding that the general pCRE-based model performed as well as a model using both the general pCREs and the shoot pCREs (AUC-ROC=0.80; **Figure 3.4B**). This further supports the notion that shoot up-regulated genes may be regulated by a global set of TFs (**Figure 3.1B**) that bind to set of general pCREs. Another surprise was that, for root up-regulated gene prediction, the models based on the root pCREs, the general pCREs, and the union of the general and the root pCREs performed similarly (**Figure 3.4A**). One potential explanation is that each model captured a distinct subset of the organ up-regulated genes. To assess the extent to which the models were predicting similar sets of genes, we examined how genes were classified when different pCRE subsets were used (see **3.5 Methods**). We found that more root specifically up-regulated genes were predicted with the root pCRE-based models (24%) compared to models using general pCREs (9%; **Figure S.3.2**).

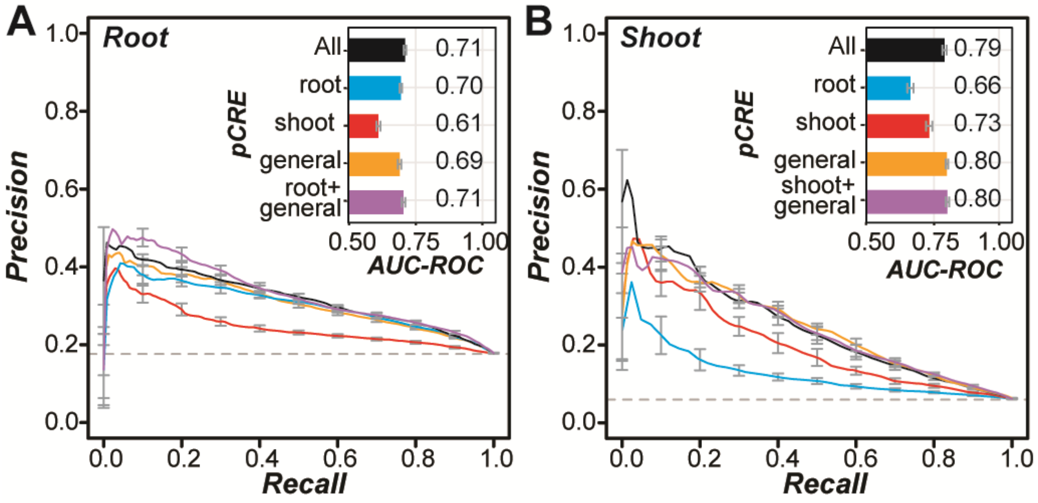


Figure 3.4 Performance of salt up-regulation prediction models using pCREs identified from co-expression clusters. (A) Precision-recall curves for models predicting root salt up-regulated genes using all pCREs (black), root pCREs (blue), shoot pCREs (red), general pCREs (orange) and, root+general pCREs (purple). The bar plot on top right indicates the corresponding AUC-ROC values of the models. Error bar is the standard error of precision values or AUC-ROCs from 10-fold cross validation. **(B)** Precision-recall curves and AUC-ROC values for models predicting shoot up-regulated genes using all pCREs (black), root pCREs (blue), shoot pCREs (red), general pCREs (orange) and, shoot+general pCREs (purple).

Taken together, these results demonstrate that the identification of the pCRE set using stress expression data can lead to improvements in modeling gene expression over known in vitro TF binding sites. This supports our hypothesis that co-expression based approaches would improve CRE discovery. We also found that salt stress up-regulated genes in the root and the shoot may be regulated by different subsets of motifs in the pCRE set. Genes up-regulated by salt stress in the root can be best predicted with a model considering both root and general pCRE sets without considering shoot pCREs. However, the shoot up-regulated genes likely are regulated primarily by general pCREs, as seen in the equivalent performance of the general pCRE model and the full pCRE model of shoot up-regulated genes.

3.3.5 Filtering pCREs based on TF binding, DNase I hypersensitivity, and conservation

We demonstrated that pCREs identified in this study can predict organ salt up-regulation. However, the large number of pCREs identified (1,894) raised the question of whether there might be motifs that were redundant or not particularly informative in the predictive model and could be filtered out. To reduce redundancy, we first removed highly similar pCRE pairs (see **3.5 Methods**). Next, we used feature selection algorithms to identify the best performing pCREs in predicting root up-regulated genes (**Figure 3.4A**) and shoot up-regulated genes (**Figure 3.4B**). Among the feature selection algorithms used, the Chi-square statistic-based approach performed the best (see **3.5 Methods**; **Supplemental Figure 3**). With a threshold chi-square statistic ≥ 10, 678 (41%) and 397 (35%) of pCREs (referred to as chi10 selected pCREs) were regarded as informative and could better predict root (AUC-ROC=0.73; **Figure 3.5A**) and shoot (AUC-ROC=0.81; **Figure 3.5B**) salt up-regulation, respectively, compared to when the full set of pCREs were used (**Figure 3.4A and B**).

To improve prediction of organ salt gene up-regulation further, we took advantage of additional regulatory information including the *in vitro* TF binding data (CIS-BP and DAP-seq), chromatin accessibility measured according to DNase I Hyper-Sensitivity (DHS) experiments [40], and the Conserved Non-coding Sequences (CNS) among Brassicaceae species [35]. Although root and shoot up-regulated gene promoters were over-represented with DHS regions (FETs, all $p < 5e-13$) and with CNSs (FETs, all p<1e-12) compared to non-responsive genes, the performance of models based on only DHS or CNS were the same as random guess (AUC-ROC ~ 0.5), suggesting additional regulatory sequence information was needed. Thus, we hypothesized that a pCRE site would be more informative in predicting gene expression if it overlapped with a potential TF binding site, a chromatin accessible region, and/or CNS.

Models based on DAP-seq filtered pCREs had similar performance to the models using original unfiltered pCREs for both root and shoot salt up-regulation (AUC-ROC=0.73-74 and 0.80-81; **Figure 3.5A and B**). Because the model performance remained the same and 9-14% of the pCRE sites were removed, it is likely that filtering based on DAP-seq data eliminated some false positive pCRE sites but also true positive sites. This is also true for filtering pCRE sites based on CIS-BP data (**Figure 3.5A**). On the other hand, filtering pCRE sites based on DHS information further decreased the performance for shoot up-regulation prediction but did not impact prediction in root (**Figure 3.5B**). Thus, pCRE sites informative for predicting shoot salt up-regulation were likely removed, potentially because chromatin accessibility can only partially predict gene expression [41]. One surprising finding was that models based on pCRE sites overlapped with CNS had the worst performance in predicting both root and shoot up-regulated genes. This is likely because the CNS were identified with stringent criteria and filtering reduced true pCRE sites. In addition, this finding suggests that there are pCRE sites that are involved in organ salt up-regulation but are not highly conserved. Taken together, the pCRE information alone already led to models with the best performance in predicting organ salt up-regulation. Additional TF binding information, DHS, and CNS either did not improve or worsen the model performance.

3.3.6 pCREs work best in combinations

To evaluate what the minimal set of pCREs was for salt up-regulation predictions, we ranked all chi10 pCREs as well as those with DAP, CIS-BP, DHS, and/or CNS evidence according to importance scores generated during machine learning runs (see **3.5 Methods**). For each chi10 pCRE, it was examined five times by either applying no filter to the sites the pCRE mapped to or by filtering based on four types of evidence (DAP, CIS-BP, DHS, or CNS), this analysis was referred to as a combined approach (**Figure 3.5**). Consistent with models with CNS filtering having the lowest performance in predicting organ salt up-regulation, we found that CNS features were the least important in predictions (**Figure 3.5C and D**). Given there were 678 root and 397 shoot chi10 pCREs, we ranked 2,712 root and 1,588 shoot pCRE-evidence pairs and identified the top 100 and 10 pCREs for root or shoot up-regulation.

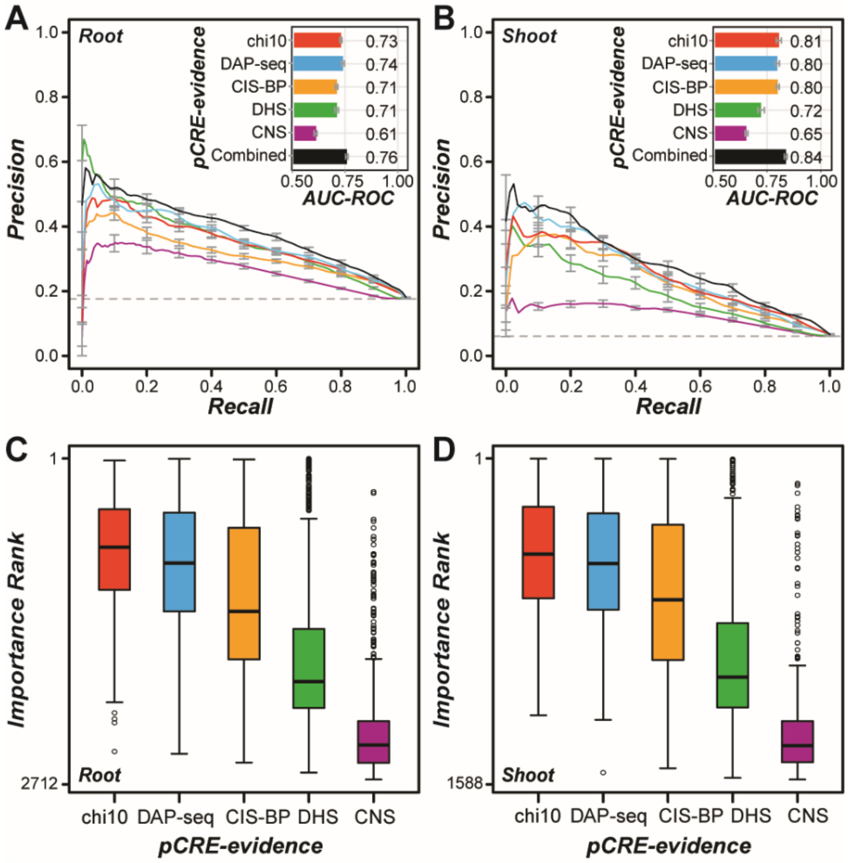


Figure 3.5 Performance of salt up-regulation prediction models using filtered pCRE sets. (A) Precision-recall curves for models predicting root salt up-regulated genes using the following pCRE sets: (1) chi10 - pCREs selected with the chi-square test feature selection method with a threshold of chi-square statistic ≥ 10 (red) (**Supplementary Figure 3E and F**), (2) DAP - chi10 selected pCREs including only pCRE sites overlapped DAP-seq peaks (blue), (3) CIS-BP - chi10 selected pCREs including only pCRE sites overlapped with CIS-BP TFBM sites (orange), (4) DHS - chi10 selected pCREs including only pCRE sites overlapped with a DNase-I Hyper-Sensitivity (DHS) peaks (green), (5) CNS - chi10 selected pCREs including only pCRE sites overlapped with a Conserved Non-coding Sequence (CNS, purple), and (6) Combined - using all information from (1)-(5) (black).

Figure 3.5 (cont'd)

(B) Precision-recall curves for models predicting shoot salt up-regulated genes using the six types of pCRE sets as in (A). **(C)** Distributions of importance ranks of all chi10 selected pCREs (chi10) or chi10 pCREs filtered based on DAP-seq, CIS-BP, DHS, or CNS data. The ranks were obtained from the model built with the Combined dataset in (A) for root. **(D)** As in (C) but based on the model built with the Combined dataset in (B) for shoot.

The models based on the top 100 pCRE-evidence pair yielded AUC-ROC=0.72 and 0.80 for predicting root and shoot up-regulation, respectively, which were comparable to the models based on all chi10 selected pCREs (red; **Figure 3.5A and B**). However, using only the top 10 pCRE-evidence pairs, the prediction performance was significantly worse (AUC-ROC=0.66 and 0.72 for root and shoot prediction, respectively). This result suggests that most important 100 pCRE-evidence pairs that included 39 and 40 pCREs for root and shoot, respectively, were sufficient for predicting organ salt up-regulated genes. This would imply that the rest of the 1,854 pCREs we identified are not informative. Alternatively, it is possible that some of these seemingly uninformative pCREs may reveal their importance only in combinations as demonstrated in studies on regulation of gene expression under stress conditions [17,20], as well as tissue specific expression [26,42]. So far, we considered many pCREs collectively in the salt up-regulation predictions but treated each pCRE individually as an independent predictor. Therefore, we asked: (1) whether pCRE combinations were important for salt up-regulated genes in the root and/or the shoots, (2) what the important pCRE combinations were, (3) what types of pCREs were involved with the combinations, and (4) if combinatorial rules important in root expression were also important for shoot expression, or *vice versa*.

To identify pCRE combinations relevant to the up-regulation of genes under salt stress, we used the Classification by Association method (CBA; see **3.5 Methods**). Due to consideration of computational complexity, we restricted our analysis to binary combinatorial rules where the presence of two pCREs predicted up-regulation (pCRE A + pCRE B → up-regulation in organ of interest). Rule sets were generated for both the root and the shoot salt up-regulated genes. As some pCREs may only be informative in combinations, we included all 1,894 pCREs without any filtering to identify combinatorial rules. This would also enable us to compare the pCREs involved in rules to the individual pCREs found to be most informative. We identified 2,826 and 351 combinatorial rules for root and shoot up-regulation. 1,086 pCREs were present in the combinatorial rules that were predictive of root up-regulation (root rules), but only 389 of them were also chi10 selected pCREs that were informative for predicting up-regulation when considered individually. Similarly, only 136 out of 427 pCREs in the shoot rules were chi10 selected. Thus, a substantial number of pCREs were informative for predicting root and shoot salt up-regulation only when considered in combination.

We also found only 12 root rules (among 2,826) had the same pCRE combinations as the shoot rules, suggesting that the great majority of the rules for one organ were specific to that organ. Most importantly, models based on only the combinatorial rule sets improved predictions for both root (AUC-ROC=0.81; **Figure 3.6A**) and shoot (AUC-ROC=0.87; **Figure 3.6B**) up-regulated genes compared to the models based on presence/absence of single pCREs (AUC-ROC=0.71 and 0.79 for root and shoot, respectively, **Figure 3.4A and B**). These results indicate the involvement of pCRE combinations to the salt up-regulation. In addition, they demonstrate that the rules capture the physical interaction between two presumed TFs binding to a pair of pCREs. Nonetheless, we found that the sites of a pair of pCREs in a rule are not significantly closer together in salt up-regulated genes compared with non-responsive genes (**Figure S.3.4A and B**). This is consistent with the finding that the distance distribution of the binding sites of interacting human TFs were not significantly different from random expectation [42]. Thus, sites of pCRE important for combinatorial regulation may not be constrained by their distances.

With the combinatorial rules, we next examined if the rules tended to be composed of a general pCRE and an organ (root or shoot) pCRE, two general pCREs, or two organ pCREs (**Figure 3.6C**). We found that there was a significant difference in the distribution of these three categories of combinatorial rules for the shoot rules (Chi-square test, *p*=6.0e-06), particularly there were more general-general pCRE combinations than expected (odds-ratio=1.5), and fewer organ-organ pCRE rules than expected (odds-ratio=0.52). This aligns with the notion that the general pCREs are more important for the regulation of shoot up-regulated genes. The root rules also had a significantly different distribution of rule types (Chi-square test, *p*=0.01), but the effect sizes were generally low (odds-ratios range 0.89-1.1). Thus, it does not appear that rules for root up-regulated genes are composed of general pCRE with a pCRE from one of the organ sets.

Taken together, our findings suggest that the organ pCREs work best in combinations and example combinatorial rules are shown in **Figure 3.6D**. The greater importance of combinatorial rules aligns well with what is already known in mammals, where individual CREs are important for expression in multiple tissues, but CRE combinations are more relevant in controlling tissue-specific expression [18,22]. Both root rules and shoot rules incorporate pCREs from the full set of organ pCREs, but there is only little overlap (0.4-3%) in the two sets of rules. This suggests that the pCREs need to be considered in combinations for better predictions of salt up-regulation.

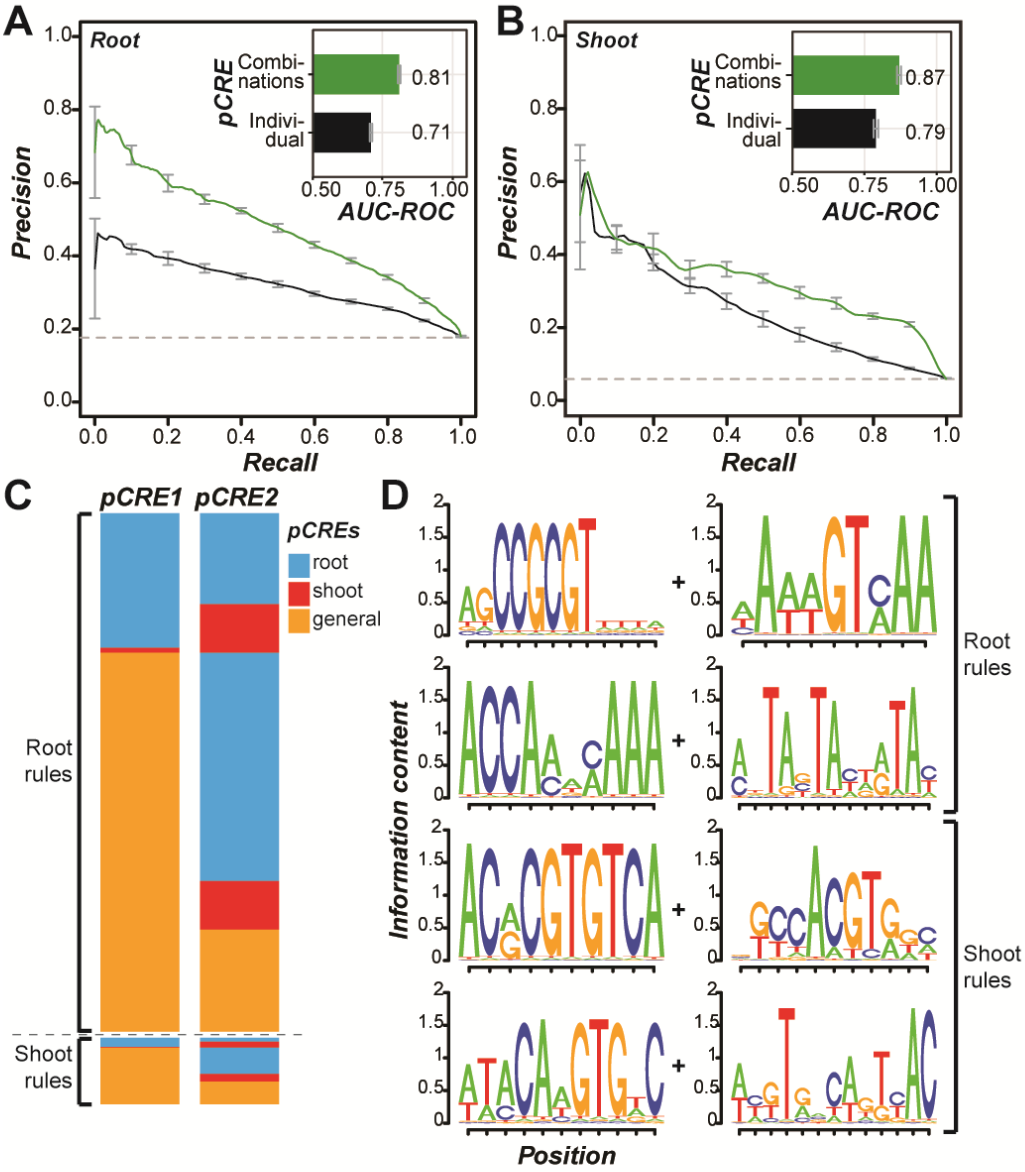


Figure 3.6 Summary of root and shoot combinatorial pCRE rules and model performance. **(A)** Precision-recall curves comparing models based on combinatorial rules (green) and the full pCRE set (black) for predicting root salt up-regulated genes. **(B)** Precision-recall curves comparing model based on combinatorial rules (green) and the full pCRE set (black) for shoot salt up-regulated genes. **(C)** Heatmap summarizing the rules identified for salt stress up-regulated genes in the root (pCREs: blue=root, red=shoot, orange=general).

3.4 Conclusions

In this study, we identified a set of 1,894 pCREs from co-expression clusters that were relevant to the up-regulation of transcript abundance under salt stress in the root and shoot of *A. thaliana*. Among them, 25% pCREs were similar to the known binding motifs of TFs from multiple families. Machine learning models for predicting salt up-regulation based on the pCRE set had significantly better performance than those based on *in vitro* binding data from two large-scale studies [22,23]. Thus, the pCREs identified likely contained *cis*-regulatory information of spatial response to salt. We also found that the salt up-regulation in the root required both a general pCRE set that was relevant to up-regulation in both root and shoot as well as a root pCRE set primarily associated with root specifically salt up-regulated genes. In contrast, the shoot salt up-regulated genes relied primarily on a general pCRE set. Considering that substantially more genes were up-regulated in the root (2,100) compared to that in the shoot (524), this difference in the composition of relevant pCREs may reflect the differences in regulatory complexity and root as the primary organ exposed to high salinity treatment. Filtering pCREs using *in vitro* TF binding data, chromatin accessibility and conservation, we found that ~40 pCREs could predict organ salt up-regulation with the same performance as the model using all pCREs. Nonetheless, the organ salt up-regulation models considering combinations of pCREs had significantly improved performance over the models considering pCREs collectively but treated each pCRE as independent predictor. Most importantly, the majority of the pCREs in the combinatorial rules were not considered important when they were treated as independent predictors and would have been false negatives in common motif finding practices.

One limitation of our study was that the pCREs were identified based on the expression data alone without knowledge of whether the sites mapped by these pCREs were actually bound by TFs. To overcome this limitation, we incorporated *in vitro* binding, chromatin accessibility, and conservation data into the model. We found that the inclusion of *in vitro* binding data led to models with the same performance as those considered only pCREs. Nonetheless, we found that the pCREs identified are complementary to *in vitro* derived TF binding information. Because the *in vitro* TF binding was an assessment of what kinds of sequences could be bound and not where the *in vivo* binding sites were in the genome, the binding data by itself was not expected to predict condition-specific expression well. Combining the pCREs identified using condition-specific data and the *in vitro* binding data, the condition-specific regulators and regulatory sequences could be pinpointed. In addition to *in vitro* TF binding data, chromatin accessibility data (DHS) was incorporated but led to either a reduced model performance or did not improve the model performance. One potential reason was because the DHS data we used were generated under conditions not related to salt stress in different developmental stages of *A. thaliana*. Finally, CNS was incorporated to filter pCRE sites but yielded models with the lowest prediction performance. One reason could be stress pCREs might have higher evolutionary rates and not well conserved. Another possible reason is that CNSs are defined in a stringent fashion or not sensitive enough in obtaining sites that are under selection but beyond the limit of detection.

Another limitation was that the spatial stress response was predicted at the organ level with limited resolution. The next logical step is to identify pCREs that can be used to predict the differential expression of genes in a cell type specific manner. In any case, our results show that co-expression based CRE identification in conjunction with machine learning-based modeling are a promising method for globally assessing spatial gene regulation in the context of stress. In addition to providing a genome-wide view of the potential *cis*-regulatory mechanisms, this approach may have possible applications in engineering plants that can respond to stresses. Use of native, tissue-specific inducible promoters to engineer plants is promising, but it is limited by the promoters that are already available in nature [43]. The methods we used here may help to identify individual and/or combinations of *cis*-regulatory sequences that can be used in synthetic promoters to drive tissue specific expression in the context of stress.

3.5 Methods

3.5.1 Expression data processing and expression data analysis

A. thaliana abiotic stress expression data for the root and shoot [11] and biotic stress data for the shoot were downloaded from the AtGenExpress website (http://www.weigelworld.org/resources/microarray/AtGenExpress/). The data came preprocessed and normalized. We calculated $\log_2$ fold changes and associated *p*-values for each stress condition and its corresponding control at each time point and each organ using limma [44] in the R environment [45]. The *p*-values were adjusted [46] to control for the False Discovery Rate. Genes were considered up-regulated if their $\log_2$ fold-change values$\geq$1 and their adjusted *p*-values$\leq$0.05. Genes up-regulated after salt treatment for three hours in the root and in the shoot were referred as the root and shoot up-regulated genes respectively. Genes were considered non-responsive if they were not significantly differentially expressed (up or down-regulated) under any stress condition at any time point in the root or the shoot. Each organ had its own set of non-responsive genes ("root non-responsive" and "shoot non-responsive"). This stringent definition of non-responsive genes was chosen to because *cis*-regulatory sequence may be relevant to regulating responses not only to salt but also to other stress conditions.

To assess the relationship of the degrees of differential expression in the root and shoot under different stress conditions, Pearson's Correlation Coefficients (PCCs) of $\log_2$ fold change values were calculated for all pairs of conditions/organ. A heatmap of the PCC values (**Figure 3.1**) was generated using the gplots package in R [47]. To identify the functional categories enriched in salt up-regulated genes (3hr) in the root, in the shoot, or in both root and shoot, each plant GO slim category (http://www.geneontology.org/ontology/subsets/goslim_plant.obo) was used to determine if it contained over/under-represented numbers of up-regulated genes in root, shoot, or both organs with Fisher's Exact Test (FET) implemented in SciPy (http://www.scipy.org/). The *p*-values were adjusted for multiple testing using the Benjamini-Hochberg method [46].

3.5.2 *In vitro* TF binding, DNase I hypersensitivity, and conserved non-coding datasets

Two sets of *in vitro* binding data were used. The first set included Position Frequency Matrices (PFMs) obtained from the CIS-BP database website [22]. These PFMs are based on either protein binding microarray data or TRANSFAC motifs [22]. The PFMs were converted to Position Weight Matrices (PWMs) adjusted for the background AT (0.33) and CG (0.17) content of *A. thaliana* genome using the TAMO package [48]. This resulted in a final set of 355 PWMs (referred to as TFBMs). To map the TFBMs, first the 1kb sequences upstream of transcriptional start sites (putative promoters) of all genes in *A. thaliana* were downloaded from The Arabidopsis Information Resource (ftp://ftp.arabidopsis.org/). The TFBMs from CIS-BP were mapped to the putative promoter sequences using Motility (http://cartwheel.caltech.edu) with a threshold p<1e-06. The second set included 344 DAP-Seq experiments testing *in vitro* bindings to naked genomic DNA in 598 TFs from the Plant Cistrome Database [25]. A DAP-seq peak (~200bp long) contained TF binding site and only peak with fraction of reads in peaks (FRiP) ≥ 5% was considered further. We identified TFBM sites and DAP-seq peaks that were over-represented in the promoters of the root up-regulated and shoot up-regulated genes by performing FET against the root non-responsive and shoot non-responsive genes, respectively.

DNase I hyper-Sensitivity (DHS) data [40] were obtained from GEO (GSE53322 and GSE53324) in form of peaks in bed format. The DHS datasets were derived from multiple developmental stages and tissues including 7-day-old dark-grown *A. thaliana* Col-0 seedlings, root, root hair cells, root non-hair cells, and seed coat. Each DHS dataset was treated as distinct features in this study for predicting salt up-regulation. *A. thaliana*-based coordinates of ~90,000 Conserved non-coding sequences (CNS) among Brassicaceae species were obtained (http://mustang.biol.mcgill.ca:8885, [35]) to assess whether CNS may be informative for assessing salt up-regulation. In addition, both DHS and CNS regions were used to filter pCRE sites to see if sites with different degrees of chromatin accessibility and conservation may contribute to salt up-regulation prediction differently.

3.5.3 Salt up-regulation pCRE identification

To identify pCREs associated with salt up-regulated genes in the root and shoot, we used a published pipeline with modifications [17]. The stress expression dataset in the form of a $\log_2$ fold-change expression matrix was used to identify co-expression clusters using iterated rounds of *k*-means clustering such that all clusters contained 60 genes or less, while clusters smaller than 10 genes were excluded. Clusters enriched in salt up-regulated genes in any time point in either roots or shoots were analyzed further for identifying 6-18bp motifs in the putative promoter regions of genes in each cluster. Six motif finding programs were used: AlignACE [49], MDScan [50], MEME [51], Motif Sampler [52], Weeder [53], and YMF [54]. In the initial motif finding step, ~300,000 motifs were identified, many of which were redundant. Two rounds of pCRE merging/enrichment testing were performed. In the first round, the ~300,000 motifs were merged if their consensus sequences shared the same IUPAC codes and/or if they were highly similar to each other (in the same cluster) based on clusters defined using Kullback-Leibler (KL) distance [17]. In the enrichment step, these merged pCREs were mapped to the 1kb promoter regions of genes in *A. thaliana* using Motility (http://cartwheel.caltech.edu), and we kept mappings with a $p < 1e\text{-}06$. The pCREs were further analyzed if their mapped sites were significantly over-represented (FET, adjusted $p \leq 0.05$) in promoters of salt up-regulated genes.

In the second round, we further merged enriched motifs based on PCC distance (1-PCC) of the motif PWMs. Using the PCC distance matrix, motifs were clustered hierarchically and distinct clusters were demarcated with a PCC distance threshold of 0.10, which was previously found to be the first percentile of PCC distances for non-redundant motifs in the JASPER CORE dataset [17]. Within each cluster, a single motif was chosen based on having the most significant degree of enrichment for genes up-regulated under salt stress in roots and/or shoots. The motifs identified from all clusters were collectively referred to as pCREs. To identify pCREs particularly relevant to root, shoot, or general salt up-regulation, a final round of FET was done to identify motifs were significantly over-represented *(*$p<0.05$) only in the root salt up-regulated genes ("root pCREs"), only in shoot salt up-regulated genes ("shoot pCREs"), and among genes up-regulated in both organs ("general pCREs"). In the end, 1,894 shoot, root, and general pCREs were identified.

3.5.4 Comparison of pCREs and TFBMs

To assess the similarity between the pCREs identified here and the known TFBMs from CIS-BP [22], the PCCs between PWMs of all pCRE-TFBM, all pCRE, and all TFBMs combinations were calculated. For each pCRE, the pCRE-TFBM combination with highest PCC was analyzed further. To assess the statistical significance of the correlation between a pCRE-

TFBM pair, a within TF family PCC distribution was established using TFBMs from each TF family. This allowed us to test whether a pCRE was more similar to a TFBM of a particular family than those between TFBMs within that same family. The PCC distribution of each TFBM family was fitted with normal or beta distribution functions based on maximum likelihood using the MASS package [55] in R. Every PCC between a pCRE and a TFBM from a particular family was compared to the cumulative density function of the fitted within family distribution to get a *p*-value. All *p*-values from the pairwise comparisons were adjusted for multiple testing within the same family [56].

To further assess which families of TFs pCREs might bind to, between family TFBM PCC distributions were generated and fitted as described above. We compared the PCC for each pCRE-TFBM pair to the between family distributions to generate a *p*-value, which were adjusted for multiple testing [56]. We set a *q*-value of 0.05 as the threshold to say that the pCRE may be bound by the same family as the TFBM. Because the TFBMs for some families were more divergent than the other (a wide range median PCCs for within family distribution; **Supplemental Figure 1**), the false negative rates (fail to assign a pCRE to certain families) varied. To assess if the pCREs were more similar to TFBMs than to random genomic sequences, 1,894 random PWMs with the same length distribution as pCREs were generated. For a random PWM of length k, 15 *k*-mers were randomly generated using the background distribution of AT-GC in *A. thaliana*, and consolidated into a PWM using the MotifTools.Motif_from_counts function in TAMO [48]. The random PWMs were then compared to TFBMs to establish the distribution of PCC to randomized PWMs.

3.5.5 Prediction of salt up-regulation using machine learning and feature selection

Our goal was to model salt up-regulation of genes in each organ as a classification problem involving two classes: salt up-regulated genes in an organ and genes that are not responsive under any stress condition. The Support Vector Machine (SVM, [57]) and Random Forest (RF, [58]) algorithms were used for classification implemented in the Waikato Environment for Knowledge Analysis (Weka [59]). To get the importance scores of each feature, RF was also used from Scikit-learn package in Python [60]. Every model in this paper had two components: 1) a set of genes, each of which is classified as up-regulated or non-responsive ("expression class") and 2) a set of *cis*-regulatory sites (CIS-BP TFBMs, DAP-seq peaks, or pCREs) and their presence/absence on the putative promoter of each gene ("promoter features"). We established machine learning models using five sets of pCRE sites including all mapped pCRE sites, as well as pCRE sites overlapped with CIS-BP TFBM sites, DAP-seq peaks, DHS regions, and CNS. In this setup, the models predict the genes from the two expression classes using the presence or absence of the promoter features. Grid-searches were used to find the best combination of the following three parameters in SVM: (1) the ratio of non-responsive to up-regulated genes, (2) the parameter of the soft margin, and (3) the gamma parameter of the Radial Basis Function (RBF) kernel. The latter two parameters are part of the SVM method itself. Similarly, grid-searches were used for RF predictions including (1) the ratio of non-responsive to up-regulated genes and (2) number of attributes. The ratio of negative to positive examples was achieved using the Weka class "weka.filters.supervised.instance.SpreadSubsample", which subsamples the non-responsive genes to achieve the desired ratio of up-regulated to non-responsive genes. We used 10-fold cross validation as implemented in Weka, and the average AUC-ROC from all 10 cross validation runs was calculated using the ROCR package [61]. RF model results were reported in this study as the performances of SVM and RF models were correlated and RF was easier to scale-up to large-datasets. The parameter combination with the maximum average AUC-ROC were taken as the best parameters for each model, and this maximum AUC-ROC is what we report for each model. Precision-recall curves were plotted using the output from the model with the maximum AUC-ROCs.

To eliminate redundant motifs, we used three univariate feature selection methods: 1) PCC-based using Caret R Package 2) Correlation Feature Selection-CFS in Weka (correlation is based on minimum description length (MDL), symmetrical uncertainty, and relief (Hall,1999)) and 3) Chi-squared test in Weka on the pCRE sets. For the PCC-based method, we calculated the PCC between each pair-wise feature (pCRE sites). For the pairs of features that have greater than a PCC of 0.5, kept only one of them. This is an arbitrary threshold; however, removing 15 -20% pCREs, did not change the AUC-ROC values of prediction models. With CFS method, we kept the default settings in Weka. Chi-squared test in Weka yields ranks for each pCRE based on the chi-square statistic. We used the chi-square statistic of 10 and 20 as thresholds to keep higher ranked pCREs.

3.5.6 Binary prediction of root and shoot up-regulated genes

While the AUC-ROC is a good measure of the overall performance of machine learning models, it does not indicate how well individual genes are predicted. Thus, it is possible that two models have similar levels of performance as measured by AUC-ROC resulting from the correct predictions of different sets of genes. To assess which genes were predicted by models based on different pCRE sets, and to see if different models correctly predict different sets of genes, the Weka program CrossValidationAddPredictions was used to identify whether a gene was correctly predicted as up-regulated or non-responsive during salt stress. This program makes a model as described above, but it keeps track of the prediction of each gene. We used the best parameter combination identified from the original grid-search as the basis for the binary prediction run. We chose the parameter combinations with the maximum average AUC-ROC. For that given run, maximum F-measure (harmonic mean of precision and recall, calculated using ROCR) was used as the threshold to create binary predictions for each gene. We also assessed the overlap of correctly predicted up-regulated genes (True Positives, "TP") based on models using different pCRE sets by looking at the percentage of the up-regulated genes correctly predicted by two different models.

3.5.7 Combinatorial motif rule discovery

To test if the combinations of specific pCREs were predictive of salt up-regulation in the root or shoot, the Classification by Association (CBA) [63] method was used to identify combinatorial rules of the form "pCRE A + pCRE B → up-regulation" were selected from the CBA output. This method is useful for identifying rules where some combinations of features are associated with a class. The features in our case were the presence or absence of pCRE pairs on a gene promoter and the class was root or shoot up-regulation. The root or shoot up-regulated and non-responsive genes were broken up into different subsamples. Each of these subsamples was run through CBA using multiple values for minimum confidence (percentage of genes where "pCRE A + pCRE B → up-regulation" out of all the instances of "pCRE A + pCRE B") and support (percentage of genes that with the rule "pCRE A + pCRE B → up-regulation"). Rules for shoot up-regulated genes were discovered using a minimum support 0.5% and a minimum confidence of 60.0%, with a non-responsive to up-regulated ratio of 2:1. We went through several rounds of CBA to discover root rules using different values of support, confidence and non-responsive to up-regulated ratios. We ended up using a minimum support of 0.1%, a minimum confidence of 60%, and subsamples with 976 non-responsive genes to 488 responsive genes, which were the same numbers of genes used to generate the shoot rules. These parameters were chosen because the rules generated gave an appreciable gain in the AUC-ROC when performing predictions. Due to the limitation of using a GUI version of CBA, we were not able to do an extensive exploration of the best CBA parameter values. Thus, it is possible that there is a more optimal parameter set that will yield a greater performance gain.

The distance between pairs of pCREs in a rule was calculated for all instances of the rules in the putative promoters. The minimal distance between the closest ends of two pCREs were determined. To determine if the minimal distances were significantly different than randomly expected, background distributions of pCREs was generated by modeling the frequency of distances between two random pCREs of the same lengths as the pCREs in the rule pair based on an earlier approach [42]. The only difference in our method was that we compared our observed distance distributions to the background distribution using a Mann-Whitney test instead of a Kolmogorov-Smirnov test, as the Mann-Whitney test can more directly test whether one distribution has higher or lower distances than the other distribution.

3.6 Acknowledgements

We thank members of the Shiu laboratory, especially Bethany Moore, Ming-Jung Liu and Melissa Lehti-Shiu, for valuable discussions and John Lloyd for help with Random Forest.

APPENDIX

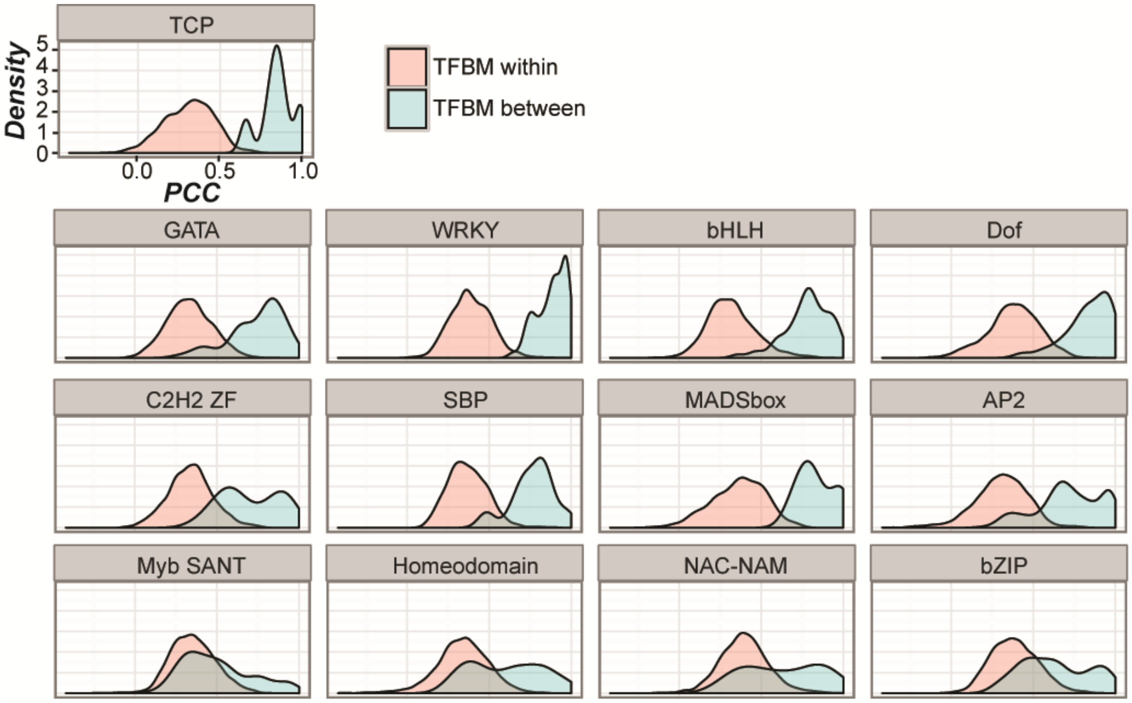


Figure S.3.1 Distributions of PCCs between CIS-BP TFBMs with and between example TF families. Distributions of PCCs between TFBMs within a TF family (red) and between TFBMs of a particular family to all other TFBMs across families (blue).

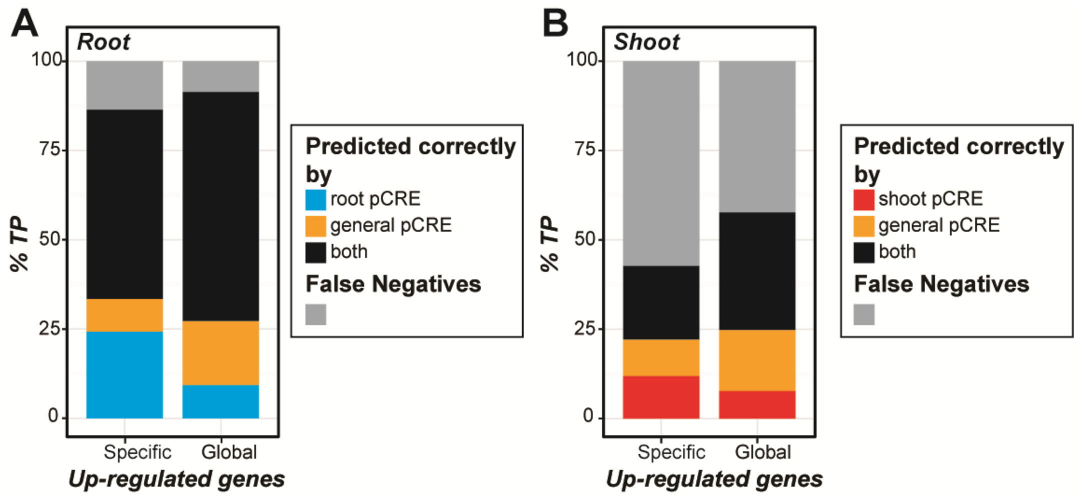


Figure S.3.2 Contribution of general, root, and shoot pCREs to the predictions of true positive genes that were globally, root-specifically, and shoot-specifically up-regulated. (A) Bar plot of % true positive root-specific or global salt up-regulated genes predicted by the root pCRE only model (blue), the general pCRE only model (orange), and both models (black). Grey: false negatives. **(B)** Bar plot of % true positive shoot-specific or global salt up-regulated genes predicted by the shoot pCRE only model (red), the general pCRE only model (orange), and both models (black). Grey: false negatives.

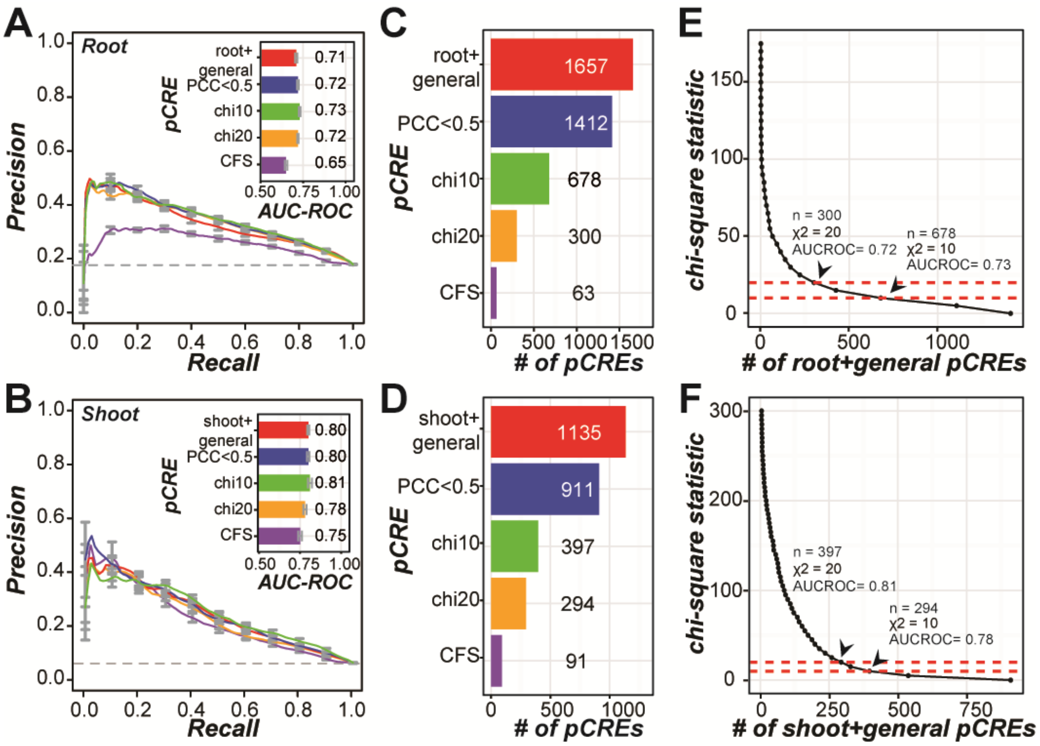


Figure S.3.3 Feature selection on pCREs and performance of RF models using selected pCREs. (A) Precision-recall curves of models predicting root up-regulated genes **(B)** Precision-recall curves of models predicting shoot up-regulated genes. Bar plots on top right of precision-recall curves indicate the AUC-ROC values for predicting root (A) and shoot (B) salt up-regulated genes. The features used are: "root+general" or "shoot+general" (red), "PCC<0.5": pCREs that have PCCs higher than 0.5 (blue), "chi10": pCREs that have chi-squared statistic higher than 10 (green). "chi20": pCREs that have chi-squared statistic higher than 20 (orange). "CFS": pCREs selected from Correlation Feature Selection (purple). **(C)** Bar plot of number of pCREs that each set of feature selection method yields for root+general pCRE set **(D)** Bar plot of number of pCREs that each set of feature selection method yields for shoot+general pCRE set. (**E**) Dot plot of chi-square statistic for root+general pCRE set. Red dashed lines: Thresholds used to select pCREs. **(F)** Dotplot of chi-square statistic for shoot+general pCRE set.

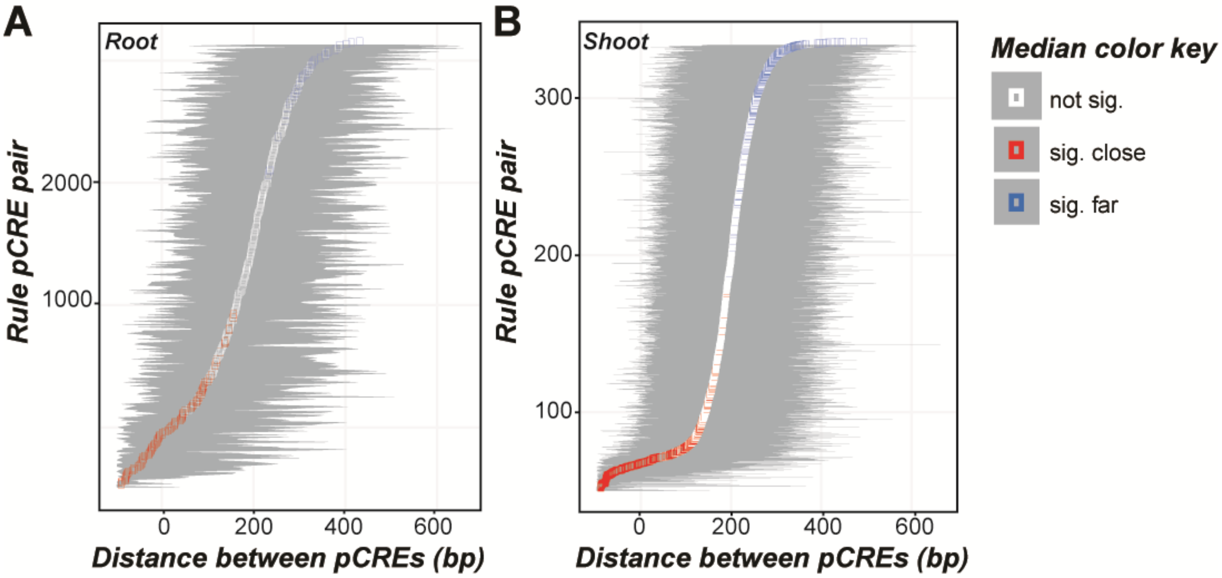


Figure S.3.4 Summary of the distance between pairs of motifs in combinatorial rules. (A) The distance between pairs of motifs in the same rule for all instances of the rule on the promoters of all genes in the *A. thaliana* genome for the root rules. Squares represent the median distance for a rule, and the edges of the ribbon represent the 25th and 75th percentile of distances. The color of each square represents the significance (Mann-Whitney, adjusted p < 0.05) of the distance distribution compared the random background distribution. White: Not significant, red: significantly closer than random, blue: significantly further than random. Rows are sorted from lowest median distance at the bottom to highest median distance at the top. **(B)** Same as (A) but for shoot rules.

REFERENCES

REFERENCES

1. Bostock RM, Pye MF, Roubtsova T V. Predisposition in Plant Disease: Exploiting the Nexus in Abiotic and Biotic Stress Perception and Response. Annu Rev Phytopathol. Annual Reviews ; 2014;52: 517–549. doi:10.1146/annurev-phyto-081211-172902

2. Rasheed S, Bashir K, Matsui A, Tanaka M, Seki M. Transcriptomic Analysis of Soil-Grown Arabidopsis thaliana Roots and Shoots in Response to a Drought Stress. Front Plant Sci. 2016;7. doi:10.3389/fpls.2016.00180

3. Cramer GR, Urano K, Delrot S, Pezzotti M, Shinozaki K. Effects of abiotic stress on plants: a systems biology perspective. BMC Plant Biol. BioMed Central; 2011;11: 163. doi:10.1186/1471-2229-11-163

4. Gargallo-Garriga A, Sardans J, Pérez-Trujillo M, Rivas-Ubach A, Oravec M, Vecerova K, et al. Opposite metabolic responses of shoots and roots to drought. Sci Rep. Nature Publishing Group; 2014;4: 6829. doi:10.1038/srep06829

5. Pierik R, Testerink C. The art of being flexible: how to escape from shade, salt, and drought. Plant Physiol. American Society of Plant Biologists; 2014;166: 5–22. doi:10.1104/pp.114.239160

6. Munns R, Tester M. Mechanisms of Salinity Tolerance. Annu Rev Plant Biol. 2008;59: 651–681. doi:10.1146/annurev.arplant.59.032607.092911

7. Munns R. Comparative physiology of salt and water stress. Plant Cell Environ. 2002;25: 239–250. doi:10.1046/j.0016-8025.2001.00808.x

8. Kreps JA, Wu Y, Chang H-S, Zhu T, Wang X, Harper JF. Transcriptome Changes for Arabidopsis in Response to Salt, Osmotic, and Cold Stress. Plant Physiol. 2002;130: 2129–2141. doi:10.1104/pp.008532

9. Dinneny JR, Long TA, Wang JY, Jung JW, Mace D, Pointer S, et al. Cell Identity Mediates the Response of Arabidopsis Roots to Abiotic Stress. Science (80-). 2008;320: 942–945. doi:10.1126/science.1153795

10. Geng Y, Wu R, Wee CW, Xie F, Wei X, Chan PMY, et al. A Spatio-Temporal Understanding of Growth Regulation during the Salt Stress Response in Arabidopsis. Plant Cell. 2013;25: 2132–2154. doi:10.1105/tpc.113.112896

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

12. Qin F, Shinozaki K, Yamaguchi-Shinozaki K. Achievements and Challenges in Understanding Plant Abiotic Stress Responses and Tolerance. Plant Cell Physiol. 2011;52: 1569–1582. doi:10.1093/pcp/pcr106

13. Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, et al. Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J. Blackwell Science Ltd; 2002;31: 279–292. doi:10.1046/j.1365-313X.2002.01359.x

14. Haberer G, Wang Y, Mayer KFX. The Non-coding Landscape of the Genome of Arabidopsis thaliana. Genetics and Genomics of the Brassicaceae. New York, NY: Springer New York; 2011. pp. 67–121. doi:10.1007/978-1-4419-7118-0_3

15. Beer MA, Tavazoie S. Predicting Gene Expression from Sequence. Cell. 2004;117: 185–198. doi:10.1016/S0092-8674(04)00304-6

16. Wang X, Haberer G, Mayer KF. Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. BMC Genomics. 2009;10: 284. doi:10.1186/1471-2164-10-284

17. Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, et al. Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2011;108: 14992–7. doi:10.1073/pnas.1103202108

18. Priest HD, Filichkin SA, Mockler TC. Cis-regulatory elements in plant cell signaling. Curr Opin Plant Biol. 2009;12: 643–649. doi:10.1016/j.pbi.2009.07.016

19. Austin RS, Hiu S, Waese J, Ierullo M, Pasha A, Wang TT, et al. New BAR tools for mining expression data and exploring Cis -elements in Arabidopsis thaliana. Plant J. 2016; doi:10.1111/tpj.13261

20. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004;431: 99–104. doi:10.1038/nature02800

21. Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci U S A. 2014;111: 2367–72. doi:10.1073/pnas.1316278111

22. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. 2014;158: 1431–1443. doi:10.1016/j.cell.2014.08.009

23. O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell. 2016;165: 1280–1292. doi:10.1016/j.cell.2016.04.038

24. Jiao Y, Lori Tausta S, Gandotra N, Sun N, Liu T, Clay NK, et al. A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. Nat Genet. 2009;41: 258–263. doi:10.1038/ng.282

25. Yáñez-Cuna JO, Kvon EZ, Stark A. Deciphering the transcriptional cis-regulatory code. Trends Genet. 2013;29: 11–22. doi:10.1016/j.tig.2012.09.007

26. Hu Z, Gallo SM. Identification of interacting transcription factors regulating tissue gene expression in human. BMC Genomics. 2010;11: 49. doi:10.1186/1471-2164-11-49

27. Zhong S, He X, Bar-Joseph Z, Harbison C, Gordon D, Lee T, et al. Predicting tissue specific transcription factor binding sites. BMC Genomics. BioMed Central; 2013;14: 796. doi:10.1186/1471-2164-14-796

28. Tsai ZT-Y, Shiu S-H, Tsai H-K. Contribution of Sequence Motif, Chromatin State, and DNA Structure Features to Predictive Models of Transcription Factor Binding in Yeast. PLoS Comput Biol. 2015;11: e1004418. doi:10.1371/journal.pcbi.1004418

29. McLeay RC, Leat CJ, Bailey TL. Tissue-specific prediction of directly regulated genes. Bioinformatics. 2011;27: 2354–60. doi:10.1093/bioinformatics/btr399

30. Zhu J-K. Salt and Drought Stress Signal Transduction in Plants. Annu Rev Plant Biol. 2002;53: 247–273. doi:10.1146/annurev.arplant.53.091401.143329

31. Golldack D, Lüking I, Yang O. Plant tolerance to drought and salinity: stress regulating transcription factors and their functional significance in the cellular transcriptional network. Plant Cell Rep. 2011;30: 1383–1391. doi:10.1007/s00299-011-1068-0

32. Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K. AP2/ERF family transcription factors in plant abiotic stress responses. Biochim Biophys Acta - Gene Regul Mech. 2012;1819: 86–96. doi:10.1016/j.bbagrm.2011.08.004

33. Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K. NAC transcription factors in plant abiotic stress responses. Biochim Biophys Acta - Gene Regul Mech. 2012;1819: 97–103. doi:10.1016/j.bbagrm.2011.10.005

34. Matiolli CC, Tomaz JP, Duarte GT, Prado FM, Bem LEV Del, Silveira AB, et al. The Arabidopsis bZIP Gene AtbZIP63 Is a Sensitive Integrator of Transient Abscisic Acid and Glucose Signals. Plant Physiol. 2011;157: 692–705. doi:10.1104/pp.111.181743

35. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;45: 891–8. doi:10.1038/ng.2684

36. Chaves MM, Flexas J, Pinheiro C. Photosynthesis under drought and salt stress: regulation

mechanisms from whole plant to cell. Ann Bot. 2009;103: 551–560. doi:10.1093/aob/mcn125

37. Ji H, Pardo JM, Batelli G, Van Oosten MJ, Bressan RA, Li X. The Salt Overly Sensitive (SOS) Pathway: Established and Emerging Roles. Mol Plant. 2013;6: 275–286. doi:10.1093/mp/sst017

38. Barah P, B N MN, Jayavelu ND, Sowdhamini R, Shameer K, Bones AM. Transcriptional regulatory networks in Arabidopsis thaliana during single and combined stresses. Nucleic Acids Res. Oxford University Press; 2016;44: 3147–64. doi:10.1093/nar/gkv1463

39. Kang J, Choi H, Im M, Kim SY. Arabidopsis Basic Leucine Zipper Proteins That Mediate Stress-Responsive Abscisic Acid Signaling. Plant Cell Online. 2002;14: 343–357. doi:10.1105/tpc.010362

40. Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, et al. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Rep. 2014;8: 2015–2030. doi:10.1016/j.celrep.2014.08.019

41. Liu M-J, Seddon AE, Tsai ZT-Y, Major IT, Floer M, Howe GA, et al. Determinants of nucleosome positioning and their influence on plant gene expression. Genome Res. 2015; gr.188680.114-. doi:10.1101/gr.188680.114

42. Yu X, Lin J, Zack DJ, Qian J. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. Nucleic Acids Res. 2006;34: 4925–4936. doi:10.1093/nar/gkl595

43. Potenza C, Aleman L, Sengupta-Gopalan C. Targeting transgene expression in research, agricultural, and environmental applications: Promoters used in plant transformation. Vitr Cell Dev Biol - Plant. 2004;40: 1–22. doi:10.1079/IVP2003477

44. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; doi:10.1093/nar/gkv007

45. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2012.

46. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57: 289–300. doi:10.2307/2346101

47. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data. 2015.

48. Gordon DB, Nekludova L, McCallum S, Fraenkel E. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs.

Bioinformatics. 2005;21: 3164–5. doi:10.1093/bioinformatics/bti481

49. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol. 1998;16: 939–945. doi:10.1038/nbt1098-939

50. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol. 2002;20: 835–839. doi:10.1038/nbt717

51. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 1994;2: 28–36.

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

53. Pavesi G, Mereghetti P, Zambelli F, Stefani M, Mauri G, Pesole G. MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. Nucleic Acids Res. 2006;34: W566–W570. doi:10.1093/nar/gkl285

54. Sinha S, Tompa M. A statistical method for finding transcription factor binding sites. Proc Int Conf Intell Syst Mol Biol. 2000;8: 344–354.

55. Venables WN & Ripley BD. Modern Applied Statistics with S. Fourth Edition. Springer, New York.2002. ISBN 0-387-95457-0

56. Storey JD. The positive false discovery rate: a bayesian interpretation and the q-value 1. Ann Stat. 2003;31: 2013–2035.

57. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20: 273–297. doi:10.1007/BF00994018

58. Breiman L. Random Forests. Mach Learn. Kluwer Academic Publishers; 2001;45: 5–32. doi:10.1023/A:1010933404324

59. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004;20: 2479–81. doi:10.1093/bioinformatics/bth261

60. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

61. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21: 3940–3941. doi:10.1093/bioinformatics/bti623

62. Hall MA. Correlation-based feature selection for machine learning. PhD Thesis. University

of Waikato.1999.

63. Ma BLWHY. Integrating classification and association rule mining. Proceedings of the 4th. 1998.

CHAPTER 4

PREDICTIVE MODELS OF CELL TYPE HIGH SALINITY RESPONSIVE GENE EXPRESSION

4.1 Abstract

How multicellular organisms respond to their environment depends on the responses of individual cell types to the surrounding fluctuations. Transcriptional reprogramming based on the environmental changes is the key to these responses. Therefore, how transcriptional reprogramming is controlled in each distinct cell type is important to decipher. However, the mechanism of cell type specific gene expression regulation, particularly towards environmental changes, is mostly unknown in plants. Here, we use *Arabidopsis thaliana* root cell type data as examples to understand the mechanism of stress responsive gene expression regulation via *cis*-regulatory elements (CREs). We used a computational approach to identify 3,095 putative CREs (pCREs) and built predictive models of high salinity responsive gene expression in six root cell types (columella, cortex, endodermis-quiescent center, epidermis, proto-phloem and stele). We found that pCREs we identified can be used to predict high salinity responsive gene expression better than random predictions. Overall, our predictive models are better at identifying non-responsive genes rather than identifying high-salinity responsive genes. Also, whole organ associated CREs yield similar prediction performance to the cell type pCREs. Therefore, the cell type pCREs identified need to be further explored to better understand the mechanism of cell type environmental responsive gene expression.

4.2 Introduction

Since the invention of the microscope, scientists have aimed to study different types of cells and the characteristics that make them different [1]. Distinguishing cell types requires studying unique functions that separate them from others. One of the crucial components of having various cell types with different functions is a result of how genes are regulated differentially and precisely controlled in response to changing environmental conditions. Based on this differential regulation, each cell type can be identified by its specific gene expression profile. However, the gene expressions, that is steady-state mRNA levels, might not reflect individual changes per cell type if the whole organ is analyzed [2]. Thus, isolating cell types is required to study cell type specific gene expression profiles. In order to isolate distinct cell types, techniques such as fluorescent activated cell sorting (FACS) and laser capture microdissection (LCM) are being used [3–5]. These techniques have been used to study *Arabidopsis thaliana* root cell types. *A.thaliana* root has been an ideal system to study cell types as the roots have a radial organization with layers of cell types (epidermis, cortex, endodermis and stele [6]) and undergo continuous development from the stem cells, meaning cells divide, expand and specialize [6]. To separate the root cell types, green fluorescent protein (GFP) lines that are specific to cell types have been developed and extensively used [7]. The ability to isolate the root cell types led the studies to investigate how individual cell types vary in gene expression and respond to environmental fluctuations.

Among the environmental fluctuations, high salt concentrations in soil impact plants adversely and result in reduction of the yield in crops [8]. How each cell type is contributing to the overall root and whole plant response to salt stress can be learned by understanding how cell types regulate salt stress responsive gene expression. Gene regulation involves multiple players, including transcription factors (TFs), cofactors, and chromatin remodeling complexes [9]. Among these players, sets of regulatory sequences that are accessible to TF binding could be used to study the differential gene expression across cell types [10]. TF binding sites can be determined in various ways such as using chromatin immunoprecipitation (ChIP) methods, array or sequencing based [11]. However, ChIP methods only cover a single TF binding and it is not feasible to cover all TFs in an organism in a single experiment. To determine TF binding sites for as many TFs as possible, *in vitro* methods such as Protein Binding Arrays (PBMs) and DNA Affinity Purification (DAP) methods have been developed [12,13]. Apart from *in vivo* and *in vitro* methods in identifying TF binding sites, computational approaches have also been successfully used [14] in hypothesizing putative *cis*-regulatory elements. These computational approaches include co-expression, phylogenetics and combination of these two [15]. For example, Haberer *et al* used PhyloCon (Phylogenetic Consensus, [16]) in combination with co-expression across 81 microarray studies in identifying candidate TF binding sites [15]. Also, Zou *et al* identified stress related putative *cis*-regulatory elements (pCREs) using co-expression across 16 stress conditions and time points, and six motif finders [17]. Given the large number of potential TF binding sites and/or pCREs that could be identified by the computational approaches (over 60,000 sites in Haberer *et al*, 1,215 pCREs in Zou *et al*), the false positive rate is likely high and additional tests are required to be confident of a handful of pCREs through experimental validation. In this regard, evaluation of pCREs with machine learning approaches is useful [17]. The machine learning approaches can take individual (and/or combinations of) pCREs as the predictors and build models for the outcome, such as cell type salt responsive gene expression. With these prediction models, it is possible to identify important pCREs that could be involved in TF binding and drive gene expression [17,18].

Regulatory mechanisms that are responsible for cell type specific responses to external factors still remain to be deciphered [19]. In this study, we aimed to investigate the *cis*-regulatory code (CRC) of salt responsive gene expression in columella (COL), cortex (COR), stele (STE), proto-phloem (PHL), epidermis (EPI) and endodermis (END) in the roots and expand the CRC of the organ salt up-regulation. Firstly, we asked whether previously identified TF binding information could predict root cell type salt up-regulation. Next, we explored to what extent the salt responsive gene expression is similar between whole root and the individual root cell types and used previously identified organ pCREs to predict root cell type salt up-regulation. We also identified pCREs that might be involved in salt up-regulation in each cell type and found there are common pCREs among cell types as well as cell type specific ones based on over-representation in the promoters of salt up-regulation genes. We built prediction models utilizing these pCREs and found that depending on the cell type, different sets of pCREs are needed to be considered for the best performing prediction models.

4.3 Results and discussion

4.3.1 Known TF binding data in predicting cell type salt responsive gene expression

To investigate to what the extent the current knowledge of TF binding data can explain salt up-regulation in the root cell types COL, COR, END, EPI, PHL, and STE [20], we built machine learning models. We used *A. thaliana* TF binding data from two large-scale *in vitro* studies, CIS-BP [12] and DAP-seq [13] that cover binding data of 758 TFs (~38% of the known *A. thaliana* TFs). Given the extensive coverage of the *in-vitro* TF binding data, we expected root cell type salt up-regulation might be explained with the known TF binding data and formed prediction models using machine learning. Using known TF binding data as predictors in Random Forest (RF), root cell type salt up-regulation predictions were better than random, and CIS-BP data performed slightly better than DAP-seq data in all six cell type predictions (AUC-ROC=0.63-0.71 and AUC-ROC=0.58-0.68 for CIS-BP and DAP-seq respectively; **Figure 4.1A**). Among these predictions, END and STE had the highest performances and they have the least and the highest number of salt up-regulated genes respectively (precision-recall curves for END and STE predictions are given in **Figure S.4.1A and B**)**.** These results suggest that the *in-vitro* TF binding data are useful in understanding root cell type up-regulation, however there is still room for improvement since the prediction performances were still lower than perfect classification (AUC-ROC ~1). Given that the CIS-BP and DAP-seq studies were not conducted on the condition that we study (high salinity), it is expected that using TF binding information that is obtained/predicted under the relevant conditions might be more informative in cell type gene expression predictive models. One such data could be the pCREs that we identified from root stress responsive gene expression.

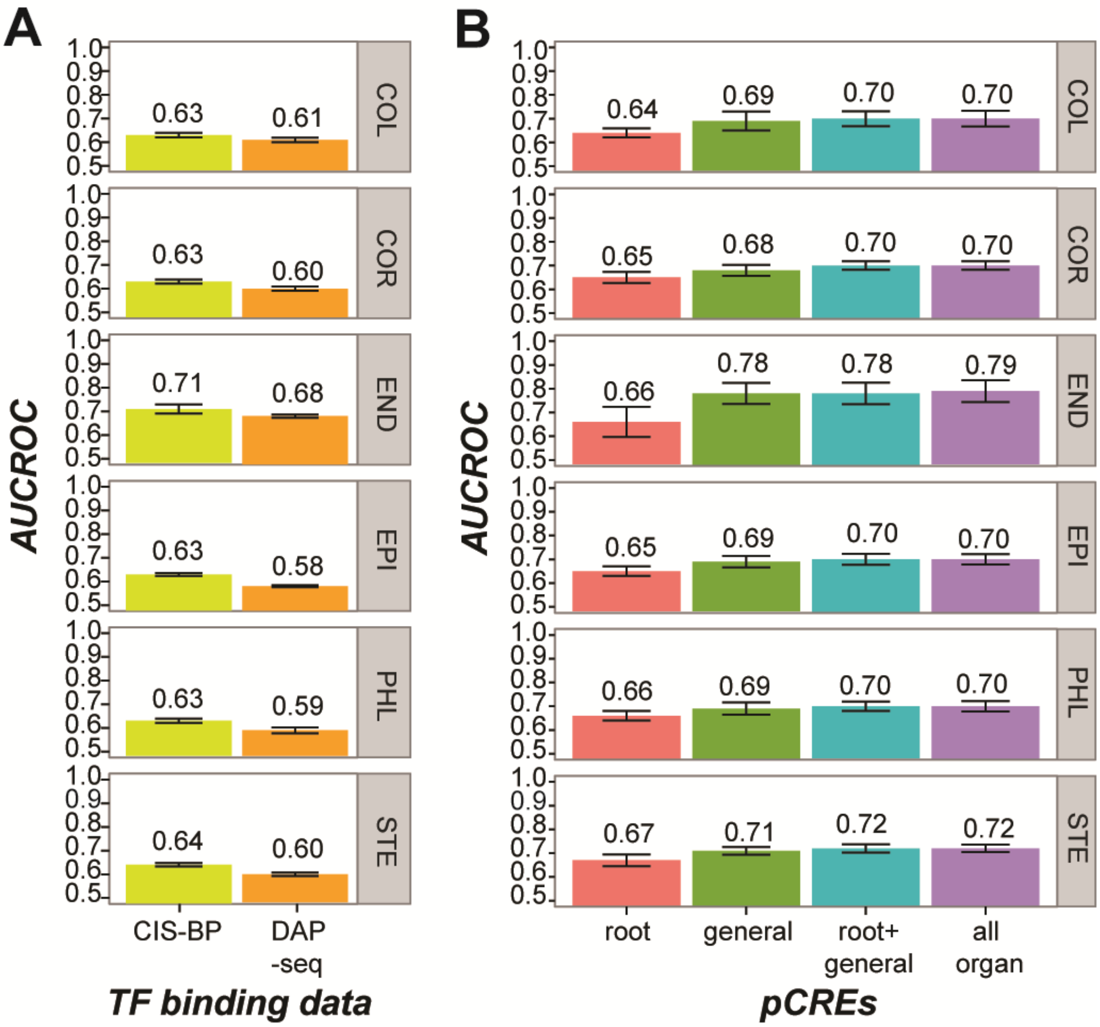


Figure 4.1 Performance of root cell type salt up-regulation prediction models using TF binding data and organ pCREs. **(A)** Bar plot of AUC-ROC values of prediction models using CIS-BP (yellow) and DAP-seq (orange) data. **(B)** Bar plot of AUC-ROC values of prediction models using organ pCREs: root (pink), general (green), union of root and general (blue), and all organ (root+general+shoot; purple) pCREs.

4.3.2 Organ pCREs in predicting cell type salt responsive gene expression

Next, to understand whether the pCREs identified from the whole organ (root and shoot) for salt stress may provide more information than *in-vitro* TF binding data in cell type salt up-regulation predictions, we used organ pCREs as the predictors in the machine learning models (**Figure 4.1B**). Our hypothesis was that the organ pCREs, particularly root pCREs, might explain cell type specific salt up-regulation because the genome-wide expression patterns of the salt-treated root cell types as well as the salt and osmotic stress-treated whole root are positively correlated (PCCs ≥ 95th percentile PCC from all pair-wise sample comparisons; **Figure 4.2**). However, we found that the root pCRE-based models did not outperform known TF binding data in predicting salt up-regulated genes in each cell type (AUC-ROC=0.64-0.67; **Figure 4.1B**). This suggests that the root pCREs are not predicting salt up-regulation to a finer resolution. Possibly with whole root expression dataset, signals from individual cell types are lost, therefore the organ pCREs do not represent the regulatory information required for the cell type salt stress responsive gene expression. Next, we used pCREs that were over-represented in both root and shoot salt up-regulated genes (referred to as general pCREs), the union of root and general pCREs, and all organ pCREs in predictions. We found that general pCREs were the best performing ones among other pCRE sets in predicting root cell type salt up-regulation (AUC-ROC=0.68-0.78; **Figure 4.1B,** Examples of the precision-recall curves are given in **Figure S.4.1C and D** for END and STE predictions). This suggests that the pCREs that are responsible for a common salt response in the roots and shoots are also able to predict cell type salt up-regulation well. It is possible that the general pCREs are representatives of common stress responsive elements regardless of the tissue. Since general pCREs likely predict common cell type salt response, the question remains what the cell type specific components are in salt response and whether the cell type specific pCREs would further improve salt up-regulation models providing a finer resolution for mechanism of salt stress responsive gene expression.

A possible way to improve salt up-regulation prediction models is to focus on the gene expression at a fine spatial resolution, namely, the gene expression of genes in individual cell types to identify cell type specific pCREs. In human studies, cell type specific CREs were identified using the gene expression data across cell types and these CREs were used to predict cell type specific gene expression [21,22]. We hypothesized that using *A.thaliana* root cell type stress expression data and identifying co-expressed gene clusters, we could find cell type pCREs that can explain cell type salt up-regulation better than the known TF binding data and whole organ pCREs.

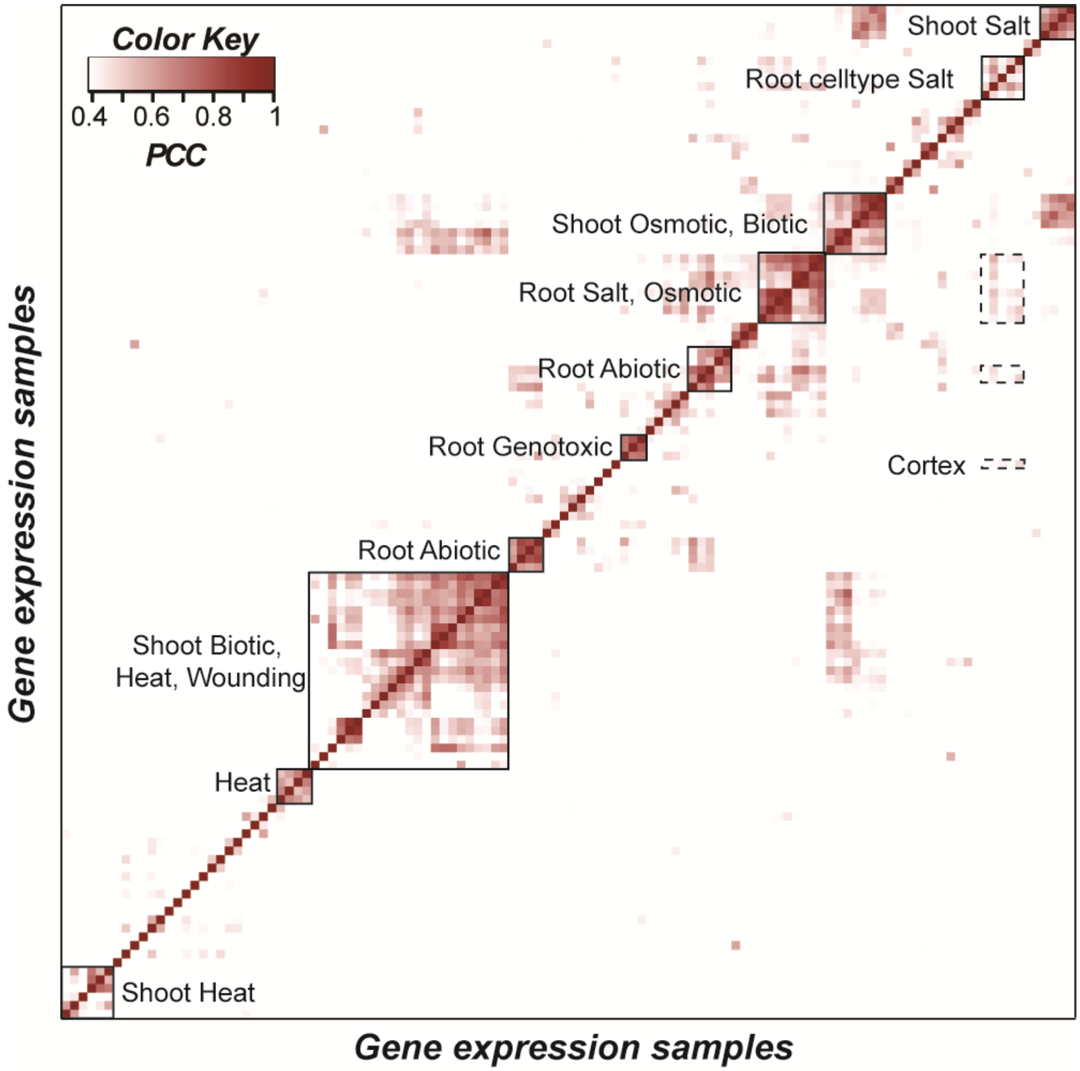


Figure 4.2 Gene expression correlation across stress datasets of root, shoot and root cell types. Heatmap of gene expression correlations using pair-wise sample comparisons. Colors represent PCC values that fall to ≥ 95th percentile of all pair-wise sample PCCs (0.42). Boxes outlined with black are the clusters of biologically relevant associations (e.g. root salt treatment samples clustering with root osmotic treatment samples). Boxes outlined with dashed black line are the correlations of 5 root cell type salt treatment samples (COL, END, EPI, PHL, STE) with root salt treatment samples, mixture of root abiotic stress samples and COR salt treatment sample respectively from top to bottom.

4.3.3 Identifying root cell type pCREs associated with salt stress

To identify root cell type pCREs that might be involved in salt stress, we used a motif finding pipeline that was developed in a previous study [17]. We incorporated root cell type differential expression in COL, COR, END, EPI, PHL, and STE [20], with the root abiotic stress data [23] to identify co-expression clusters over-represented with salt stress up-regulated genes in different root cell types. After selecting the these clusters, we identified pCREs from the promoters of the co-expressed genes. According to the enrichment of pCRE sites in the promoters of the genes in these clusters, we classified pCREs into subsets. 583 pCREs were classified as root cell type general pCREs, 734-2828 pCREs were considered specific to a particular cell type or found to be over-represented in multiple cell type salt up-regulated genes, and 6-360 pCREs were considered specific to only one cell type (**Figure 4.3A**). Because we hypothesized that we would find new cell type pCREs, we expected to have motifs that were distinct from organ pCREs. Additionally, we asked to what extent each pCRE subset has similar pCREs and expected that within the same set, pCREs might be more similar than those across different sets. To address these questions, we calculated the average PCC among the pCREs within each pCRE subset. We found that within a pCRE set, we did not have the most similar pCREs. For example, cell type general pCREs had the highest average correlation (PCC=0.78), but EPI and STE specific pCREs had average PCCs ≤ 0.4. We also found that the organ pCRE set and cell type pCRE set did not have high similarity across subsets (PCC=0.37-0.47) supporting the notion that the pCREs we identified from genes up-regulated in different cell type were distinct from the set of pCREs identified using data from whole root and shoot (**Figure 4.3B**).

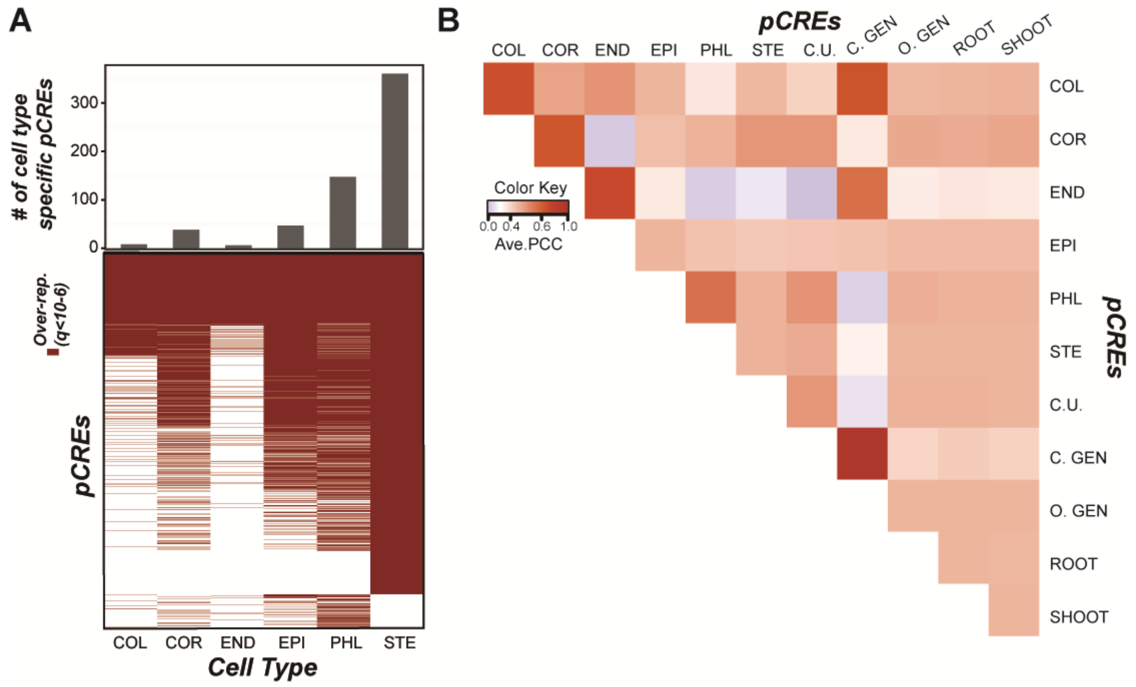


Figure 4.3 Cell type pCREs: Classification and similarity among pCRE sets. (A) Top panel: Bar plot of number of cell type specific pCREs. Bottom panel: Heatmap of over-represented pCREs. Each row is a pCRE and red color is for over-representation of that pCRE in the cell type salt up-regulated genes. **(B)** Heatmap of similarity among pCRE sets. Similarity is calculated as PCC between Position Weight Matrices (PWMs). COL, COR, END, EPI, PHL and STE pCREs are cell type specific pCREs. C.U. is cell type union, referring to pCREs that belong to more than one cell type pCRE set. C. GEN is the cell type general pCREs that are over-represented in all cell type salt up-regulated genes.

4.3.4 Cell type pCREs in predicting salt up-regulation

We have shown that cell type pCREs were distinct from the organ pCRE set, suggesting that the novel motifs may be important in driving salt up-regulation among root cell types in addition to the general organ pCREs. We used the cell type pCREs in predicting salt up-regulation of each cell type. We found that the general cell type pCRE were better at predicting END salt up-regulation than END pCREs (AUC-ROC=0.75 vs 0.55; **Figure 4.4A**; **Figure S.4.2A**). This trend was also observed with COL, COR and EPI salt up-regulation predictions; the general cell type pCREs performed better than the respective cell type pCREs (**Figure 4.4A**). On the other hand, STE salt up-regulation seemed to be driven by STE pCREs as the prediction performance was better when STE pCREs were use compared to using general cell type pCREs (AUCROC=0.69 vs 0.63 respectively; **Figure 4.4A**; **Figure S.4.2B**). This was also the case for PHL salt up-regulation predictions. For all the predictions, the union of cell type and general pCREs gave the best performances, reflecting that both general stress response pCREs and cell type specific pCREs might be responsible for the cell type responses. Also, depending on the cell type the major driving force in salt up-regulation could be either general or cell type pCREs. Overall, the performance of cell type pCREs was similar to the organ pCREs in predicting up-regulation in various cell types (**Figure 4.1B**; **Figure 4.4A**) and combining these pCRE sets (organ pCREs+cell type pCREs) did not further improve prediction model performances (**Figure 4.4B**). Even though we have seen similar performances of the salt up-regulation prediction using the organ pCRE set and cell type pCRE set, potentially these sets could predict different sets of genes to be salt up-regulated. To test this, we compared the sets of genes predicted by models based on the organ pCRE set and cell type pCREs, focusing on the STE up-regulated genes as an example. We found that ~50% of the true positives predictions were the same from the two models (**Figure S.4.3A**). Furthermore, 10%

of the salt responsive genes were correctly predicted by the organ pCREs only and 14% were predicted correctly by the cell type pCREs only. This result implies that we could predict an additional set of the salt responsive genes using cell type pCREs. However, it is interesting to note that organ pCREs also predict a subset of STE up-regulated genes that the cell type pCREs could not classify correctly. For these sets of genes that were predicted by only organ pCREs, by only cell type pCREs and by both, we asked whether the levels of salt up-regulation were different (e.g. cell type pCREs predicting genes that have a higher salt responsive gene expression-higher fold change- in STE). Our expectation was that there might be differences in the salt up-regulation levels between the genes predicted only by certain pCRE sets. However, the gene sets predicted (10%, 14%, 50% of STE) did not differ in fold change in salt responsive gene expression (**Figure S.4.3B**). This suggests that there may be another layer in which these genes differ in so that different sets of pCREs are informative in predicting them. Overall, we could predict salt up-regulation to a finer resolution in *A. thaliana*, yet it remains to be investigated whether the cell type pCRE set could be improved in explaining the specific cell type response.

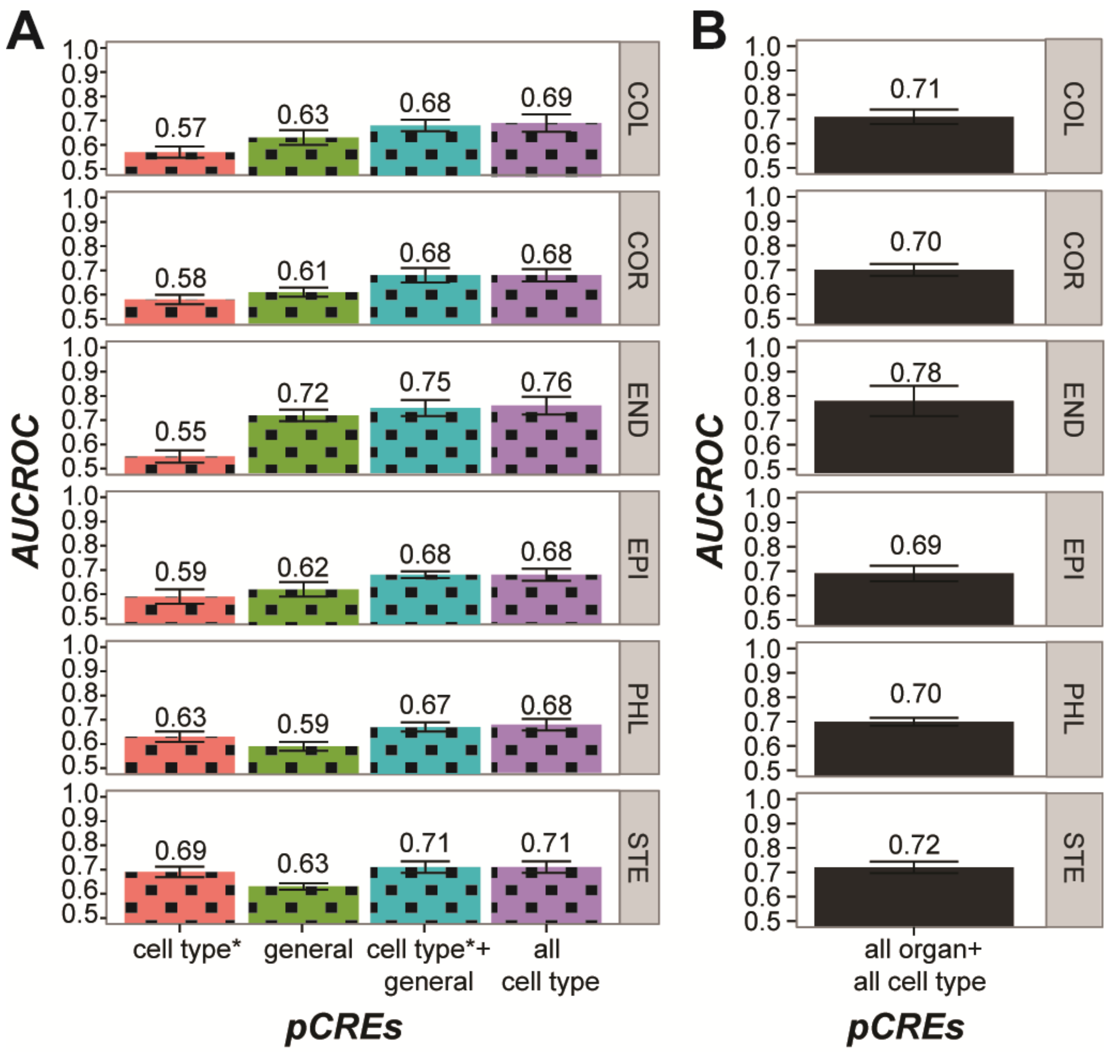


Figure 4.4 Performance of cell type salt up-regulation prediction models using cell type pCREs. **(A)** Bar plot of AUC-ROC values of prediction models using cell type (pink), general (green), union of cell type and general (blue), and all cell type (purple) pCREs. *: Each cell type has a different set of pCREs over-represented in the salt up-regulated gene sets **(B)** Bar plot of AUC-ROC values of prediction models using the union of organ and cell type pCREs.

4.4 Conclusions

With the abundance of transcriptomics studies, it is possible to study the gene regulatory networks of different tissues and cell types under changing environmental conditions. In this study, we identified pCREs that might be involved in the cell type specific salt stress responsive gene expression, particularly salt up-regulation. We found evidence that CRC regulating genes across the whole organs might be partially responsible for regulating the root cell type expression based on the prediction performance of machine learning models with organ general pCREs as predictors (**Figure 4.1B**). However, we could not rule out the possibility that cell type specific pCREs were important in predicting cell type salt up-regulation. Particularly, for STE, we identified pCREs that were able to predict STE up-regulated genes well and 14% of the STE up-regulated genes were only predicted with cell type pCREs (**Figure S.4.3A**). This shows that even though organ and cell type pCREs lead to similar performances in predicting salt up-regulation across cell types, each pCRE set can explain a different portion of genes that are up-regulated. Overall, we identified pCREs and built computational models that can explain cell type salt up-regulation.

To further improve the results of this study and overcome the limitations, the following efforts can be made. Firstly, the cell type gene expression data used here only consist of one time point (salt treatment for 1h). It was the only dataset available at the time of data processing stage of this project. Since then, more studies were conducted related to how root cell types respond to environmental fluctuations and salt stress [24]. There is a dataset consisting of 6 time points (1h, 3h, 8h, 20h, 32h, 48h) of salt treatment across 4 root cell types (COL, COR, EPI and STE) [24]. We analyzed this dataset with the expectation that including more time points would further improve the co-expression clustering and subsequent pCRE identification. However, we found that 1h salt treatments of the same cell types from two different datasets [20,24] did not yield similar expression (based on expression sample clustering similar to **Figure 4.2**). This is a common issue known as the batch effect [25] in gene expression studies and needs further analyses. For future work, it would be necessary to evaluate the time scale data more extensively to remove the batch effects and incorporate this dataset into the pCRE identification.

Apart from the expression data used, the approach in analyzing the potential regulatory motifs could also be improved. We can further expand on root cell type salt up-regulation by identifying *k*-mers. *K*-mers are consensus sequences that are of length *k* and over-represented in a given set of sequences (e.g promoters of co-expressed genes). Studies [9,26] found that *k*-mers are informative in predicting gene expression. In this study, we used six motif finders in finding PWMs, however using *k*-mers could be informative in predicting root cell type gene expression and in our initial analyses we found promising results. Combinatorial rules are another aspect that could be constructed for improving cell type salt up-regulation predictions. As organization of CREs is the key in the transcription factor complexes to form and drive gene expression [27], considering pCRE pairs might be more important in regulating cell type salt up-regulation compared to individual pCREs, as we have seen in organ salt responsive predictive models (Chapter 3).

Apart from the points discussed in improving cell type salt up-regulation predictions, other questions to expand on the CRC involved in regulating cell type processes are: (1) What are the pCREs involved in the root cell identity and how different are pCREs involved in cell identity vs. cell response? (2) What are the pCREs involved in spatial salt down-regulation and what are the differences between CRC of up-regulation and. down-regulation? Through answering these questions, it might be possible to get a more detailed genome-wide view of the spatial and conditional CRC in plants.

4.5 Methods

4.5.1 Gene expression dataset

Root cell type expression data was downloaded as CEL files from GEO (GSE7641, [20]). This expression dataset consists of control and salt stress conditions (150mM NaCl treatment for 1h) applied to columella (COL), cortex (COR), endodermis+quiescent center (END), epidermis (EPI), proto-phloem (PHL) and stele (STE). The CEL files were pre-processed and quantile normalized using the Bioconductor affy package in R environment [28]. Log_2 fold changes and the *p*-values were calculated for salt stress and the corresponding control for each cell type using the limma package [29]. *p*-values were corrected for multiple testing using false discovery rate [30]. In addition to root cell type expression data, root abiotic stress data from AtGenExpress (http://www.weigelworld.org/resources/microarray/AtGenExpress/) were used and the data were processed in a previous study [17]. Up-regulation in each cell type was defined for the genes with $\log_2$ fold-change values≥1 and their adjusted *p*-values≤0.05. Non-responsive genes were defined as neither up or down-regulated under any stress at any time point in the root or in any cell type sample.

4.5.2 Co-expression analyses

To find co-expressed gene clusters, root stress expression dataset from AtGenExpress was combined with root cell type salt stress expression dataset and 20,060 genes in the expression dataset were clustered into co-expression clusters using *c*-means [31] in the R environment. Among the resulting clusters, clusters that had greater than 10 genes and less than 60 genes were selected for further analyses. Clusters that were larger than 60 genes were clustered further and clusters containing less than 10 genes were excluded from the analyses. This range of number of genes in a cluster was required for efficiently running the motif finders [17]. Overall 538 clusters were obtained. Fishers exact test was used to select the clusters among the 538 that were over-represented with cell type salt up-regulated genes (*q*-value ≤0.05) [32]. This analysis was repeated for each cell type separately.

4.5.3 Cell type pCRE identification

To identify cell type pCREs, a previously established pipeline using six motif finders was used [17]. Motifs were found in the putative promoters (-1kb) of the genes that were in the co-expression clusters over-represented with each cell type salt up-regulated genes. Overall, 7,417 and 3,095 pCREs were identified for cell type salt up-regulation and these pCREs were over-represented in the cell type salt up-regulated genes (for at least one cell type) with an over-representation *q*-value of 0.05 and 10^{-6} respectively. Among the 3,095 pCREs, 978 of them were over-represented in the promoters of COL salt up-regulated genes compared to the rest of the genome; 1728 in COR, 734 in END, 2,340 in EPI and 2,828 in STE. 583 pCREs were commonly over-represented in all root cell type salt up-regulated genes and were referred as “general” pCREs. Note that cell type up-regulated genes include cell type specific up-regulated genes as well as the genes that were up-regulated in more than one cell type. To assess similarity between pairs of pCREs, PCCs were calculated using the Position Weight Matrices (PWMs) of each pCRE.

4.5.4 Predictive models

In obtaining predictive models, Support Vector Machine (SVM, [33]) and Random Forest (RF, [34]) were used in classifying root cell type up-regulated genes and non-responsive genes in the Waikato Environment for Knowledge Analysis (Weka, [35]). To find optimal parameters for each classification, grid-searches were done with the following in SVM: (1) the ratio of non-responsive to up-regulated genes, (2) the parameter of the soft margin, and (3) the gamma parameter of the Radial Basis Function (RBF) kernel; in RF: (1) the ratio of non-responsive to up-regulated genes and (2) number of attributes to use in trees. 10-fold cross validation was performed in prediction models. Two approaches were used to evaluate the prediction performance. (1) Area Under Curve-Receiver Operating Characteristic (AUC-ROC) measure, where a perfect model would have AUC-ROC=1 and random predictions would lead to AUC-ROC=0.5. (2) Precision-recall curve, where precision is the ratio of true positive predictions to overall predicted as positive and recall is the ratio of true positive predictions to total number of positive class (salt upregulated genes). Better models would have precision-recall curves tending more towards the upper-right corner of the graph and random predictions would be no better than the background.

APPENDIX

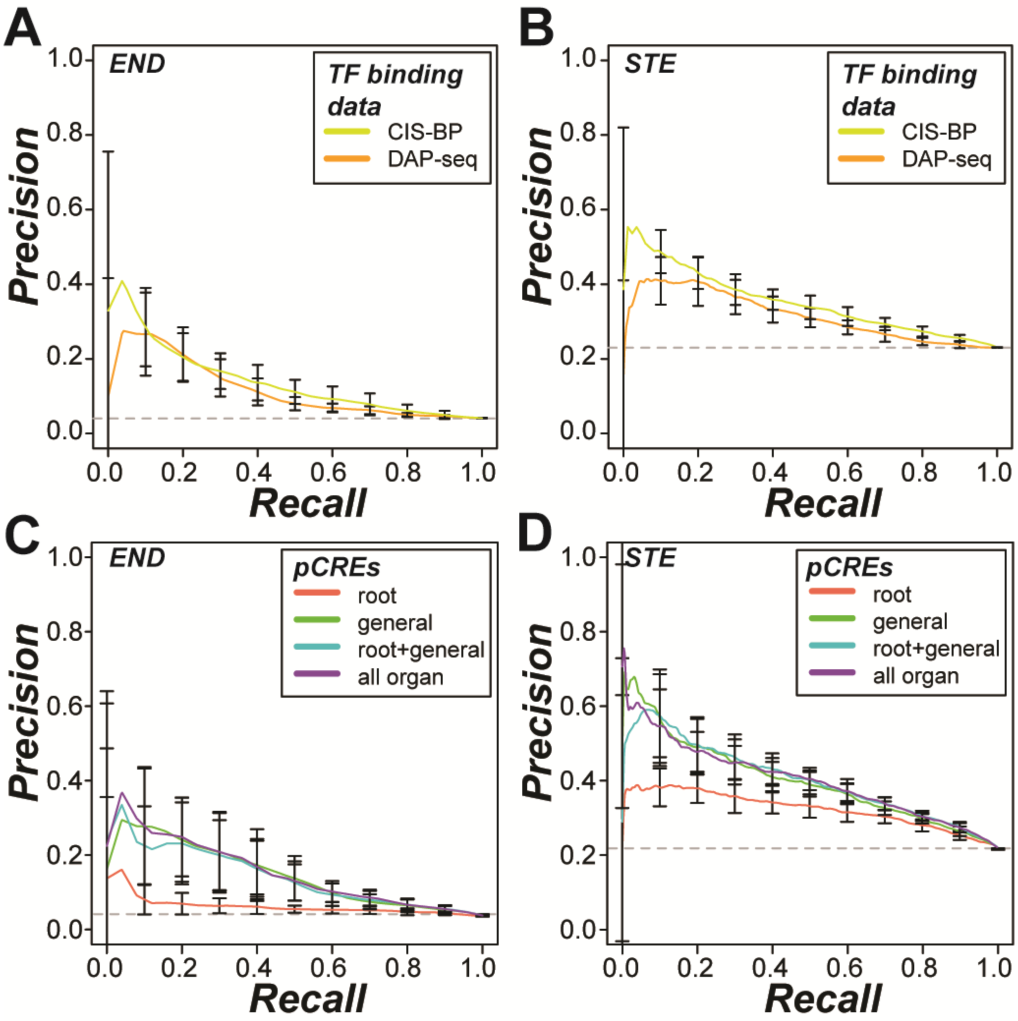


Figure S.4.1 Precision/recall of END and STE salt up-regulation prediction models using TF binding data and organ pCREs. **(A)** Precision/recall curves of END salt up-regulation models using CIS-BP (yellow) and DAP-seq data (orange). **(B)** Same as (A) but for STE salt up-regulation. **(C)** Precision/recall curves of END salt up-regulation models using root (pink), general (green), union of root and general (blue), and all organ (root+general+shoot; purple) pCREs. **(D)** Same as (C) but for STE salt up-regulation.

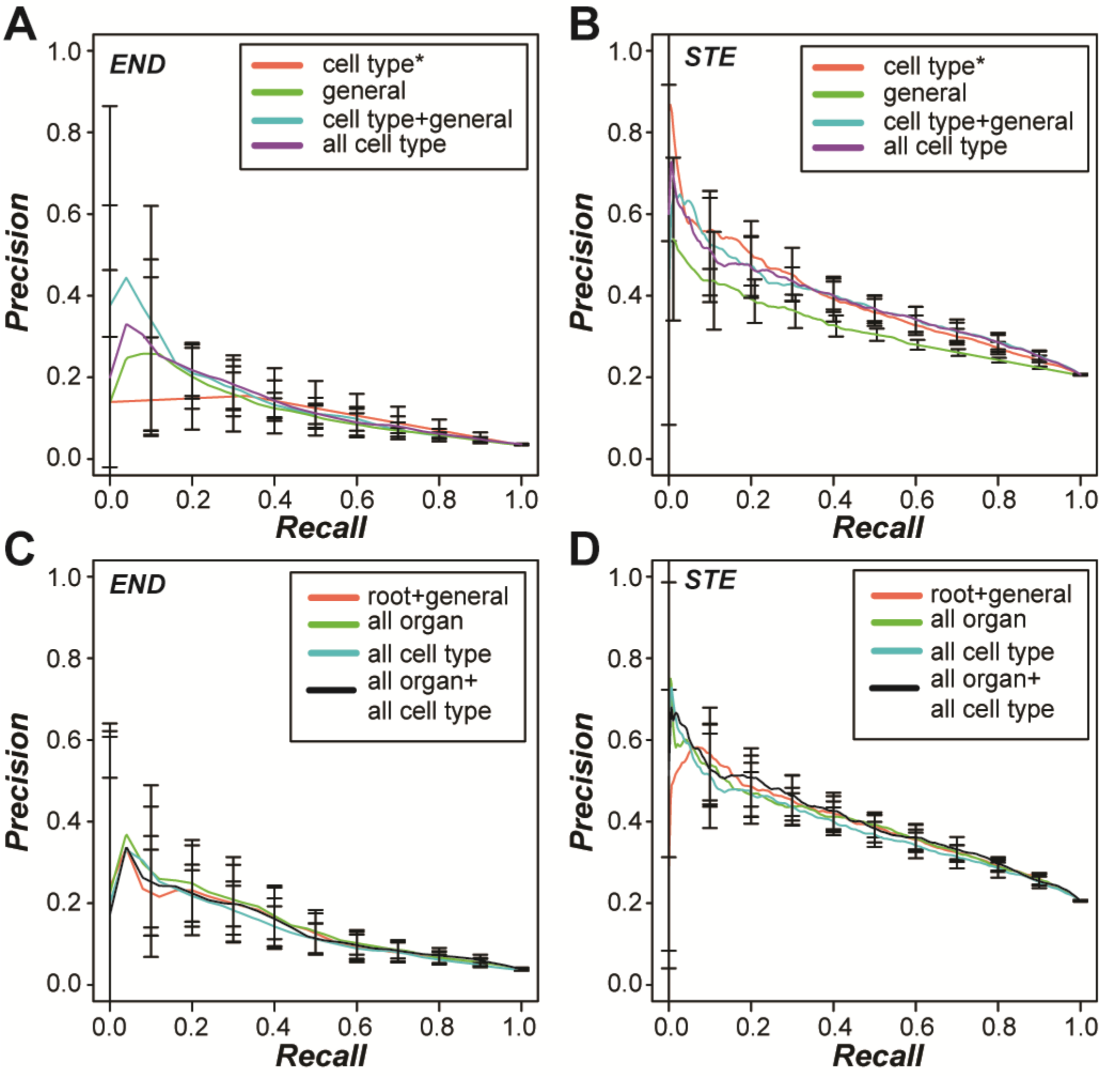


Figure S.4.2 Precision/recall of END and STE salt up-regulation prediction models using cell type pCREs and union of cell type+organ pCREs. **(A)** Precision/recall curves of END salt up-regulation models using cell type (pink), general (green), union of cell type and general (blue), and all cell type (purple) pCREs. **(B)** Same as (A) but for STE salt up-regulation. *: depends on the predicted cell type, * in (A) refers to END pCREs, in (B) refers to STE pCREs. **(C)** Precision/recall curves of END salt up-regulation models using **(D)** Same as (A) but for STE salt up-regulation.

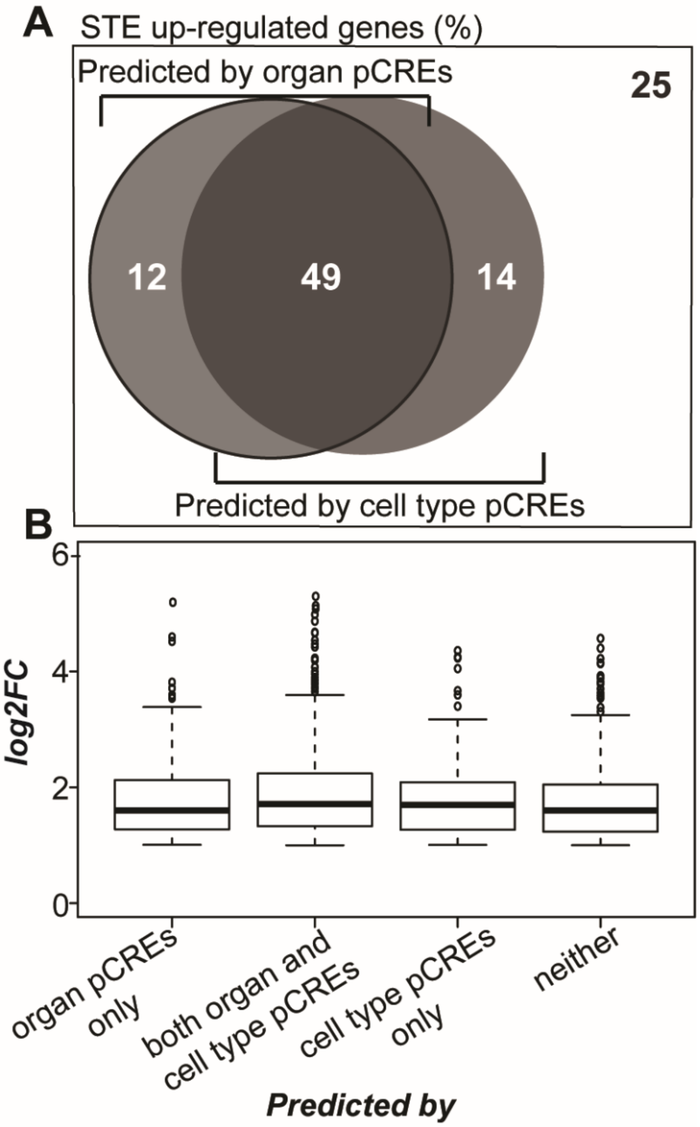


Figure S.4.3 True positive predictions from models using organ and cell type pCREs. (A) Venn-diagram of percentage true positive predictions using organ pCREs and cell type pCREs. STE up-regulated genes were divided into 4 classes: 12% that were correctly predicted by only organ pCREs, 49% that were correctly predicted by both organ and pCREs, 14% that were correctly predicted by only cell type pCREs and 25% that could not be predicted correctly by either pCRE sets **(B)** Boxplot of $\log_2$ fold change expression values of each class of STE up-regulated genes described in (A).

REFERENCES

REFERENCES

1. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res. Cold Spring Harbor Laboratory Press; 2015;25: 1491–1498. doi:10.1101/gr.190595.115

2. Benfey PN, Bennett M, Schiefelbein J. Getting to the root of plant biology: impact of the Arabidopsis genome sequence on root research. Plant J. Blackwell Publishing Ltd; 2010;61: 992–1000. doi:10.1111/j.1365-313X.2010.04129.x

3. Neira M, Azen E. Gene discovery with laser capture microscopy. Methods Enzymol. 2002;356: 282–9. Available: http://www.ncbi.nlm.nih.gov/pubmed/12418206

4. Bryant Z, Subrahmanyan L, Tworoger M, LaTray L, Liu CR, Li MJ, et al. Characterization of differentially expressed genes in purified Drosophila follicle cells: toward a general strategy for cell type-specific developmental analysis. Proc Natl Acad Sci U S A. 1999;96: 5559–64.

5. Southall TD, Gold KS, Egger B, Davidson CM, Caygill EE, Marshall OJ, et al. Cell-Type-Specific Profiling of Gene Expression and Chromatin Binding without Cell Isolation: Assaying RNA Pol II Occupancy in Neural Stem Cells. Dev Cell. Elsevier; 2013;26: 101–112. doi:10.1016/j.devcel.2013.05.020

6. Benfey PN, Schiefelbein JW. Getting to the root of plant development: the genetics of Arabidopsis root formation. Trends Genet. 1994;10: 84–8. Available: http://www.ncbi.nlm.nih.gov/pubmed/8178369

7. Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, et al. A gene expression map of the Arabidopsis root. Science. American Association for the Advancement of Science; 2003;302: 1956–60. doi:10.1126/science.1090022

8. Hirt H, Shinozaki K. Plant responses to abiotic stress. Springer; 2004.

9. Narlikar L, Ovcharenko I. Identifying regulatory elements in eukaryotic genomes. Brief Funct Genomic Proteomic. Oxford University Press; 2009;8: 215–30. doi:10.1093/bfgp/elp014

10. Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, et al. Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. Cell. Elsevier; 2013;154: 888–903. doi:10.1016/j.cell.2013.07.020

11. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet. 2012;13: 840–52. doi:10.1038/nrg3306

12. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell.

2014;158: 1431–1443. doi:10.1016/j.cell.2014.08.009

13. O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell. 2016;165: 1280–1292. doi:10.1016/j.cell.2016.04.038

14. Koryachko A, Matthiadis A, Ducoste JJ, Tuck J, Long TA, Williams C. Computational approaches to identify regulators of plant stress response using high-throughput gene expression data. Curr Plant Biol. 2015;3: 20–29. doi:10.1016/j.cpb.2015.04.001

15. Haberer G, Mader MT, Kosarev P, Spannagl M, Yang L, Mayer KFX. Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea. Plant Physiol. American Society of Plant Biologists; 2006;142: 1589–602. doi:10.1104/pp.106.085639

16. Wang T, Stormo GD. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics. Oxford University Press; 2003;19: 2369–80. doi:10.1093/BIOINFORMATICS/BTG329

17. Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, et al. Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proc Natl Acad Sci U S A. 2011;108: 14992–7. doi:10.1073/pnas.1103202108

18. Li Y, Chen C yu, Kaye AM, Wasserman WW. The identification of cis-regulatory elements: A review from a machine learning perspective. BioSystems. 2015. pp. 6–17. doi:10.1016/j.biosystems.2015.10.002

19. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol. Nature Research; 2015;16: 144–154. doi:10.1038/nrm3949

20. Dinneny JR, Long TA, Wang JY, Jung JW, Mace D, Pointer S, et al. Cell Identity Mediates the Response of Arabidopsis Roots to Abiotic Stress. Science (80-). 2008;320: 942–945. doi:10.1126/science.1153795

21. Natarajan A, Yardımcı GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type–specific gene expression from regions of open chromatin. Genome Res. 2012;22: 1711–1722. doi:10.1101/gr.135129.111

22. Chen C, Zhang S, Zhang X-S. Discovery of cell-type specific regulatory elements in the human genome using differential chromatin modification analysis. Nucleic Acids Res. 2013;41: 9230–9242. doi:10.1093/nar/gkt712

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

24. Geng Y, Wu R, Wee CW, Xie F, Wei X, Chan PMY, et al. A Spatio-Temporal Understanding of Growth Regulation during the Salt Stress Response in Arabidopsis. Plant Cell. 2013;25: 2132–2154. doi:10.1105/tpc.113.112896

25. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. Oxford University Press; 2007;8: 118–27. doi:10.1093/biostatistics/kxj037

26. Lapidot M, Michal L, Mizrahi-Man O, Pilpel Y. Functional characterization of variations on regulatory motifs. PLoS Genet. 2008;4: e1000018. doi:10.1371/journal.pgen.1000018

27. Ezer D, Zabet NR, Adryan B. Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. Comput Struct Biotechnol J. Research Network of Computational and Structural Biotechnology; 2014;10: 63–9. doi:10.1016/j.csbj.2014.07.005

28. Bioconductor. [cited 12 Apr 2016]. Available: http://www.bioconductor.org/

29. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; doi:10.1093/nar/gkv007

30. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57: 289–300. doi:10.2307/2346101

31. Pal NR, Bezdek JC, Hathaway RJ. Sequential Competitive Learning and the Fuzzy c-Means Clustering Algorithms. Neural Networks. Elsevier Science Ltd.; 1996;9: 787–796. doi:10.1016/0893-6080(95)00094-1

32. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100: 9440–5. doi:10.1073/pnas.1530509100

33. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20: 273–297. doi:10.1007/BF00994018

34. Breiman L. Random Forests. Mach Learn. Kluwer Academic Publishers; 2001;45: 5–32. doi:10.1023/A:1010933404324

35. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004;20: 2479–81. doi:10.1093/bioinformatics/bth261

CHAPTER 5

CONCLUDING REMARKS

To decipher functional associations among *Arabidopsis thaliana* genes, and regulatory information of stress responsive gene expression, I used computational approaches to analyze publicly available global gene expression data. In Chapter 2, I evaluated the utilities and limitations of using gene expression data in hypothesizing gene function. I found that based on the individual metabolic pathways, the extent of gene co-expression differs and 5%-53% of pathways form significant gene co-expression groups with similarity measure and expression dataset used impacting this percentage. I also evaluated the commonly-used clustering algorithms with different parameters and showed that focusing on only one algorithm led to information loss in relating previously unknown genes to known pathways. In validating the co-expression clusters obtained, I used an independent phenomics dataset to confirm functional associations to leucine degradation pathway. These analyses serve as an outline of the best practices of using gene co-expression in functional inference and will be an important resource for studies that aim to utilize gene co-expression analyses.

One potential improvement to the research described in Chapter 2 can be including gene expression datasets generated by different technologies. I used microarray data available for *A. thaliana* to ask to what extent pathway genes co-express and what the impact of dataset on the degree of pathway gene co-expression is. Currently, 48,501 gene expression datasets in the public data repository Gene Expression Omnibus (GEO) are microarray-based and 11,338 datasets are based on RNA-sequencing (RNA-seq) experiments (as of January 19th, 2017, [1]). Even though there are more microarray datasets compared to RNA-seq ones, the advantage of RNA-seq is that genes currently not present as probes on microarrays can be studied. It is shown that 6,953 *A. thaliana* genes are structurally annotated but are not represented on probes of Affymetrix ATH1 microarray platform, and 52% of these genes not on the array do not have functional annotation information [2]. In this case, RNA-seq data can be helpful to uncover co-expression relationships of genes absent on microarrays. In addition, RNA-seq allows more information to be extracted, such as transcript isoforms or expressed intergenic regions, that can be included in co-expression studies. For example, co-splicing networks have been constructed using RNA-seq data and correlations in the isoform ratios across different genes have been calculated for functional associations [3]. These additional associations using RNA-seq data can complement the associations obtained from using microarray data alone.

Another future direction would be to use machine learning models to predict pathway genes using gene co-expression features (e.g. expression coherence, cluster membership). In my study, I used unsupervised, clustering methods to associate unknown genes to known pathways. This goal can also be achieved in a supervised fashion by classification of pathway genes against genes that do not belong to any pathways. This classification can yield models that later be used on the rest of the genome to identify genes/genomic regions that show resemblance to genes from specific pathways. These models could also show what the most important co-expression features (i.e. EC and cluster membership) are in identifying genes that belong to specific pathways. In addition to gene co-expression, other levels of genome-wide information can be used in making functional associations. For example, protein-protein interactions can be another level of information to along with gene co-expression. Although there are data integration approaches using multiple biological data types to build gene functional networks [4,5], datasets apart from gene expression still do not have extensive temporal, spatial and conditional information. With accumulation of such data, it will be possible to investigate the utilities and limitations of using different data types for gene functional inference.

Another future direction in utilizing gene co-expression is new pathway discovery. In my study, I focused on existing literature of metabolic pathways and other biological processes. However, co-expression clusters might only contain unknown genes and these co-expressed genes could belong to a process that previously have not been studied. Using biotic stress treatment gene expression datasets, new signaling pathways have been identified from co-expression modules [6]. In the same study, known TF binding motifs such as G-box, W-box, and MYB motifs were also mapped to genes in the same co-expression clusters. Through combination of motif and pathway over-representation analyses on co-expression clusters, novel pathways as well as the potential regulatory signatures responsible for pathway gene co-expression can be identified. Similar analyses can be applied to the remaining co-expression clusters (366) obtained in my research. The co-expression clusters can also be useful in computational prediction of *cis*-regulatory elements (CREs) driving pathway gene co-expression. There are still many CREs to identify that are responsible for pathway gene expression patterns, considering I could not find over-represented known TF binding motifs on the promoters of high EC pathway genes. Therefore, additional CREs can to be identified using co-expression clusters over-represented with pathway genes to complement the TF binding motif information obtained from large scale in-vitro studies [7,8].

In research described in Chapters 3 and 4, I identified putative CREs (pCREs) that are likely involved in spatial salt responsive gene expression. These pCREs were further used in predictive models of organ and cell type salt up-regulation and contribute towards knowledge of salt stress gene regulation. In addition to pCREs, I also incorporated known TF binding sites, chromatin accessibility information, and conservation across species into machine learning models to predict organ salt up-regulation and to identify the most important pCREs responsible for salt up-regulation in root and shoots. In Chapter 4, the predictive models were expanded to the cell type level and I identified pCREs that are likely involved in root cell type specific salt up-regulation. Chapters 3 and 4 revealed pCREs that could be involved in spatial high salinity response and the predictive models generated

For future research, additional genomic regions (in addition to promoters including introns, genes) could be used to identify pCREs and information on protein-protein interactions between TFs can be considered. Although most (86% [9]) TF binding sites are found near transcription start sites, there are also regulatory elements that are known to be present in the introns as well as the genic regions [10]. Therefore, to identify pCREs, other genomic regions should also be considered. Apart from the CREs, the characteristics of the TFs that bind to these CREs also needs to be considered. For example, TFs may not function by themselves and require additional activators to initiate binding to CREs. TFs bind to cofactors or other TFs to regulate gene expression. In this respect, one limitation of our study was that our predictive models did not take TF-cofactor and TF-TF interactions into account. It is possible that some TFs only drive gene expression after these interactions are present and we would miss the information from these interactions by only focusing on the co-expressed genes and the DNA sequences. Coupling information from gene co-expression with TF-protein interactions can improve the models of organ salt up-regulation. However, conditional protein-protein interaction data, particularly under salt stress, need to be generated.

Prediction of cell type specific salt up-regulation was a challenge in this study. Cell-type data provides a higher resolution information compared to that from heterogeneous cell types within an organ. However, the current cell-type study still focus a group of cells that may be heterogenous. Recently, studies focus on subpopulations of cells sharing a common gene-expression profile with the help of single-cell sequencing studies [11]. As more data are available from single cell sequencing studies, it will be possible to utilize the single cell gene expression data to identify CREs driving individual cell gene expression. Apart from cell type predictions, prediction of down-regulations (both organ and cell type) remains a challenge. Even though predictions of organ salt up-regulation were satisfactory using computationally identified pCREs as predictors, performance of down-regulation predictions was similar to random guessing. This suggests that down-regulation is more complex than up-regulation and considering pCREs is not enough. Potentially, post-transcriptional mechanism like RNA turnover can be taken into account in modeling down-regulations. Hence, improving predictions of down-regulations will require additional data types such as RNA approach to equilibrium sequencing (RATE-seq, [12]) to be generated across multiple experimental conditions.

Overall, in my dissertation I used data-driven approaches towards learning about gene regulation and function. Even in the genomes that are well-studied, there are regions that need to be functionally annotated. The methodology and results of this thesis are applicable to annotating functional regions including genes and regulatory elements.

REFERENCES

REFERENCES

1. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. Oxford University Press; 2002;30: 207–10. doi:10.1093/NAR/30.1.207

2. Giorgi FM, Del Fabbro C, Licausi F. Comparative study of RNA-seq- and Microarray-derived coexpression networks in Arabidopsis thaliana. Bioinformatics. 2013;29: 717–724. doi:10.1093/bioinformatics/btt053

3. Aghamirzaie D, Collakova E, Li S, Grene R, Carvalho R, Feijão C, et al. CoSpliceNet: a framework for co-splicing network inference from transcriptomics data. BMC Genomics. 2016;17: 845. doi:10.1186/s12864-016-3172-6

4. Lee T, Yang S, Kim E, Ko Y, Hwang S, Shin J, et al. AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. Nucleic Acids Res. 2015;43: D996-1002. doi:10.1093/nar/gku1053

5. Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, et al. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell. 2011;23: 895–910. doi:10.1105/tpc.111.083667

6. Ma S, Shah S, Bohnert HJ, Snyder M, Dinesh-Kumar SP. Incorporating Motif Analysis into Gene Co-expression Networks Reveals Novel Modular Expression Pattern and New Signaling Pathways. Copenhaver GP, editor. PLoS Genet. Public Library of Science; 2013;9: e1003840. doi:10.1371/journal.pgen.1003840

7. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. 2014;158: 1431–1443. doi:10.1016/j.cell.2014.08.009

8. O’Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell. 2016;165: 1280–1292. doi:10.1016/j.cell.2016.04.038

9. Yu C-P, Lin J-J, Li W-H, Koudritsky M, Domany E, Lin Z, et al. Positional distribution of transcription factor binding sites in Arabidopsis thaliana. Sci Rep. Nature Publishing Group; 2016;6: 25164. doi:10.1038/srep25164

10. Riechmann JL. Transcriptional Regulation : a Genomic Overview. Arabidopsis Book. 2002;1: e0085. doi:10.1199/tab.0085

11. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. Nature Research; 2015;33: 155–

160. doi:10.1038/nbt.3102

12. Neymotin B, Athanasiadou R, Gresham D. Determination of in vivo RNA kinetics using RATE-seq. RNA. Cold Spring Harbor Laboratory Press; 2014;20: 1645–52. doi:10.1261/rna.045104.114

13. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. Nature Research; 2015;13: 165–170. doi:10.1038/nmeth.3688