Identification of common and unique stress responsive genes of *Arabidopsis thaliana* under different abiotic stress through RNA-Seq meta-analysis

Shamima Akter

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

MASTER OF SCIENCE

in

Crop and Soil Environmental Sciences

Song Li, Chair

M. A Saghai Maroof

Bo Zhang

December 19, 2017

Blacksburg, VA

Keywords: Abiotic stress, Reactive oxygen species, RNA-Seq, RNA-Seq pipeline, Gene Omnibus Series (GSE), Differentially Expressed Genes (DEG), Jaccard similarity index, Gene Ontology (GO)

Identification of common and unique stress responsive genes of *Arabidopsis thaliana* under different abiotic stress through RNA-Seq meta-analysis

Shamima Akter

ABSTRACT

Abiotic stress is a major constraint for crop productivity worldwide. To better understand the common biological mechanisms of abiotic stress responses in plants, we performed meta-analysis of 652 samples of RNA sequencing (RNA-Seq) data from 43 published abiotic stress experiments in *Arabidopsis thaliana*. These samples were categorized into eight different abiotic stresses including drought, heat, cold, salt, light and wounding. We developed a multi-step computational pipeline, which performs data downloading, preprocessing, read mapping, read counting and differential expression analyses for RNA-Seq data. We found that 5729 and 5062 genes are induced or repressed by only one type of abiotic stresses. There are only 18 and 12 genes that are induced or repressed by all stresses. The commonly induced genes are related to gene expression regulation by stress hormone abscisic acid. The commonly repressed genes are related to reduced growth and chloroplast activities. We compared stress responsive genes between any two types of stresses and found that heat and cold regulate similar set of genes. We also found that high light affects different set of genes than blue light and red light. Interestingly, ABA regulated genes are different from those regulated by other stresses. Finally, we found that membrane related genes are repressed by ABA, heat, cold and wounding but are up regulated by blue light and red light. The results from this work will be used to further characterize the gene regulatory networks underlying stress responsive genes in plants.

Identification of common and unique stress responsive genes of *Arabidopsis thaliana* under different abiotic stress through RNA-Seq meta-analysis

Shamima Akter

GENERAL AUDIENCE ABSTRACT

Abiotic stress is a major constraint for crop productivity worldwide. To better understand the common biological mechanisms of abiotic stress responses in plants, we performed analysis of 652 samples of RNA sequencing data from 43 published abiotic stress experiments in *Arabidopsis thaliana*. These samples were collected from eight different abiotic stresses including drought, heat, cold, salt, light and wounding. We identified genes that were induced or repressed by each of these stresses. We found that 5729 and 5062 genes are induced or repressed by only one type of abiotic stresses. There are only 18 and 12 genes that are induced or repressed by all stresses. The commonly induced genes are related to gene expression regulation by stress hormone. The commonly repressed genes are related to reduced growth. We compared stress responsive genes between any two types of stresses and found that heat and cold regulate similar set of genes. We also found that high light affects different set of genes than blue light and red light. Finally, we found that membrane related genes are repressed by stress hormone, heat, cold and wounding but are up regulated by blue light and red light. The results from this work will be used to further characterize the gene regulations underlying stress responsive genes in plants

DEDICATION

My Mother

(Mrs. Rowshan Akter)

A strong, beloved, and gentle soul who taught and encouraged me in hard work and so that much could be done with little effort.

My Father

(Md. Fazlul Haque Sarker)

For earning an honest living for us and for supporting and encouraging me to believe in myself.

My Husband

(Manik Ahmed)

His relentless trying for my future and always encouraging me to do toughest jobs on the earth.

My daughters

(Mumtahina S. Ahmed & Mahdita S. Ahmed)

If they were not with me I could not have write a single word, probably I could not even take breath.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and thanks to Dr. Song Li for giving me the opportunity to work in genomics and bioinformatics area which got my interest when I rotated through in his lab. I am very grateful to him as he provided directions in every way to improve myself in research field. I believe for his guidance, advices and suggestions I could be able to finish my journey in this new area successfully and now at the verge of getting my degree that will help me to move forward with new vision and ambitions. He taught me to think critically about my research. I would also like to thank my committee members Dr. M. A Saghai Maroof and Dr. Bo Zhang for their guidance, support and time.

I am thankful and grateful to my lab members Jiyoung Lee who always gave me the critical and realistic suggestions to handle the situations and problems. Several times I took suggestions and assistance from her to work in the project. Her advices, providing solutions are remarkable remembrance for me. I would like to thank Alex Qi Song who is my lab member and I met with him when I did not know anything about the coding. I got help from him. He helped me in this project by writing a very critical script to get important results.

I am also grateful to Dr. Zhang and her lab members. She gave me the opportunity to work in her project and to work in her lab for GWAS project. I want to thank Ph.D student Matthew Colson in Dr. Zhang's lab as he helped me to organize some works for sugar extraction when I was in very hardship for my daughter's illness. I would like to thank Hazem Sharaf who is my friend and classmate. He helped me several times to trouble shoot and solve problem during building the pipeline.

I am grateful to my husband Manik Ahmed for his continuous support and encouragement in every step of my study. He taught me to move forward and not to be hopeless but try. I am ever grateful to my daughters that they allowed me to study despite they need me most.

TABLE OF CONTENTS

TITLE i

ABSTRACT.....i

DEDICATION.....iv

ACKNOWLEDGEMENTS.....v

TABLE OF CONTENTS.....vi

LIST OF FIGURES viii

LIST OF TABLES.....x

CHAPTER 1.....1

LITERATURE REVIEW 1

1.1 ABIOTIC STRESS AND AGRICULTURE PRODUCTIVITY 2

1.2 MECHANISMS OF STRESS TOLERANCE: GENE EXPRESSION CHANGE 3

1.3 RNA-SEQ.....6

1.4 RNA-SEQ PIPELINE 7

1.5 SOFTWARE TOOLS OF THE RNA-SEQ PIPELINE.....7

1.5.1 READ ALIGNMENT: STAR.....7

1.5.2 READ COUNTING: FeatureCounts 8

1.5.3 DIFFERENTIAL EXPRESSION: DESeq2.....8

1.5.4 DIFFERENTIAL EXPRESSION: edgeR.....9

1.6 REFERENCES.....10

CHAPTER 2.....15

RNA-SEQ ANALYSIS OF PUBLISHED RNA-SEQ DATA FOR *ARABIDOPSIS THALIANA* UNDER DIFFERENT ABIOTIC STRESSES.....15

2.1 INTRODUCTION.....16

2.2 OBJECTIVES 17

2.3 METHODS.....18

2.3.1 PROCESS OF RNA-Seq ANALYSIS 18

2.3.2 PIPELINE OVERVIEW.....19

2.3.3 FLOW CHART OF THE PIPELINE 20

2.3.4 SCRIPTS OF THE PIPELINE 21

2.3.5 SOFTWARES OF THE PIPELINE.....22

2.3.6 DATA SETS 25

2.4 RESULTS.....26

2.5 DISCUSSION 41

2.6 CONCLUSIONS.....47

2.7 FUTURE DIRECTIONS 47

2.8 REFERENCES.....49

3.0 APPENDIXES 58

3.1 APPENDIX 1: SCRIPT FOR MAPPING READS ON GENOME 58

3.3 APPENDIX 3: SCRIPT FOR MERGING READS AND FPKM.....62

3.4 APPENDIX 4: RSCRIPT FOR DIFFERENTIAL EXPRESSION ANALYSIS 66

LIST OF FIGURES

Figure 2.1 Process of RNA-Seq analysis. Three steps in RNA-Seq analysis: 1) Experimental design 2) Sequencing 3) Data analysis in high performance computing system. 19

Figure 2.2 Flowchart of the RNA-Seq pipeline. There are seven steps in this pipeline. In each step, software names were denoted by orange text. In every step one input file is needed which is represented by blue text whereas the output file is denoted by green text. The output file of each step is the input for the next step. 21

Figure 2.3 Summary of 43 experiments that have been analyzed through the pipeline. 27

Figure 2.4 Comparison between number of up and down regulated genes in number of stress. X axis represents number of stresses and y axis represents number of genes. Blue colored bar denotes down-regulated and orange colored bar denotes up-regulated genes. This bar plot showed decreasing pattern of number of genes with increased number of stresses. 31

Figure 2.5 The Jaccard similarity index for induced genes for each stress. Jaccard index is defined as Jaccard(A,B)=|A∩B|/|A∪B|. A represents genes in one stress and B represents genes in another stress. Dendrogram is on the left and upper side of the heatmap. Stresses are in both right and lower side of the dendrogram. The upper corner scale denotes the color for similarity. Red color is highest similarity, blue color indicates no similarity and white color denotes moderate similarity. Three groups include group 1: heat, cold; group2: ABA, salt and high light and group 3: blue light, red light and wounding. 37

Figure 2.6 Jaccard similarity index for repressed genes for each stress. Dendrogram is in the left and upper side of the histogram. Name of the stresses are in both right and lower side of the dendrogram. The upper corner scale denotes the measurement of similarity: red color for highest similarity, blue for no similarity and white color denotes moderate similarity. Group1: heat, cold; Group2: salt and high light. Group 3: blue light and red light. 38

Figure 2.7 A schematic model of abiotic stress response based on genes found in our analysis. Transcription of many genes results in different mechanisms to stress tolerance. In this study of abiotic stress data analysis of *Arabidopsis thaliana,* ABF3/RD26 encodes ABREB that induced genes (GBSS1 and AT1G02660) that positively regulate energy production. P5CS1 was induced by stress and has ROS scavenging activity. EXPA16 has cell expansion or cell

wall modification activity. SRI kinases work in ABA signal transduction pathway. Stress also inhibit CAB3, a positive regulation of chloroplast activity and RGF9, a positive regulator of cell proliferation. This model is drawn using BEACON pathway editor (Elmarakeby et al., 2017). 44

LIST OF TABLES

Table 2.1 This table summarizes the 43 experiments of published data for *Arabidopsis thaliana* have been collected from GEO database. In all these data, there are 3 time series experiments, 652 SRRs with 260 conditions, 8 stress types and 19 tissue types. 25

Table 2.2 Eight stresses: blue light, high light, red light, cold, heat, salt, ABA and wounding from 10 experiments were selected. These experiments were selected because they have properly designed biological replicates. One experiment (GSE72806) has combination of salt and heat stresses. One experiment has three combined stresses including heat, cold and wounding. Two separated experiments (GSE63406 and GSE67332) tested the cold stress. ..26

Table 2.3 Summary of read mapping results from STAR software. There were 652 samples used in this analysis. 28

Table 2.4 Summary of reads counted by FeatureCounts software. There were 644 samples used in this analysis. 28

Table 2.5 Summary from the DESeq2 run for six selcted GSE experiments. 29

Table 2.6 Functional annotation of common stress responsive genes obtained using Thalemine tool in Araport https://apps.araport.org/thalemine/.....34

Table 2.7 The functional annotation of the common repressed genes using Thalemine tool in Araport https://apps.araport.org/thalemine/. Highlighted genes are known to be involved in plant stress responses and are discussed in main text. 34

Table 2.8 The detailed functions of the common induced and repressed genes. The genes in the light blue area are the induced genes and the genes in the light orange are repressed genes. 35

Table 2.9.1 Summary for gene ontology of unique up and down regulated genes in agriGO GO analysis tool.4 GO terms are significant with very low FDR. Genes are in these GO term the genes have Nutrient reservoir, TF, TF binding activity. 40

Table 2.9.2 One single GO term GO:0044425 is significant for the stresses: ABA, cold, heat, wounding with repressed genes and blue light and red light induced genes with function of

membrane part. GO annotation was done through AGRiGO go analysis tool. http://systemsbiology.cau.edu.cn/agriGOv2/.....40

CHAPTER 1

LITERATURE REVIEW

1.1 ABIOTIC STRESS AND AGRICULTURE PRODUCTIVITY[1]

Plants are living in complex environmental conditions and constantly experience abiotic stresses such as high light intensity, ultraviolet light, high and low temperatures, drought, high salinity, heavy metal, flooding and wounding (Hirayama and Shinozaki, 2010). According to the reports of the Intergovernmental Panel of Climate Change (http://www.ipcc.ch), these abiotic stresses will become more prevalent and the intensity of these stresses will increase tremendously near future.

Since the beginning of 1980s, global temperatures have increased by ~0.4° C annually and some regions experienced more increase than other regions (Lobell and Field, 2007). One direct consequence of the increasing temperature is drought stress. Drought severely reduces crop yield, because the physiology, growth, development, and productivity of crops are largely affected by the limited water availability. High salinity in soils is another common stress found across the world; approximately 7% of global land surfaces and 5% of world's cultivated land are affected by salt stress (Flowers and Yeo, 1995). Heat stress is a type of abiotic stress that can negatively impact crop productivities. For example, in 2003, the crop production was reduced by 30% by a heat wave in Europe.

Plants are sessile organisms that cannot survive without adaptation to environmental changes; abiotic stress responses are important for plant survival and reproduction (Hirayama and Shinozaki, 2010). Therefore, one of the most important topics in plant research is to

[1] Authors: Shamima Akter, Dr. Song Li. Author contributions: Shamima Akter performed literature review and prepared the first draft of this chapter. Shamima Akter and Dr. Song Li both edited this chapter.

understand the molecular mechanisms of abiotic stress responses. Molecular biology have made major progresses in this research field in the past decades: many genes involved in abiotic stress responses have been isolated and functional characterized through basic research using transgenic plants (Oh, 2005; Kikuchi et al., 2015; Wang et al., 2016). A large amount of functional data were generated to gain better insight of gene functions related to plant abiotic stress responses (Shinozaki and Yamaguchi-Shinozaki, 2007; Yoshida et al., 2014; Roy, 2016). In 2000, the genome of *Arabidopsis thaliana* were sequenced and the genomic sequences provided a major advancement for new researches in plant abiotic stress response (Arabidopsis Genome Initiative, 2000). Using the genomic sequences, gene families related to abiotic stresses have been identified and their molecular sequences can be used to study the function and evolution of abiotic stress response genes. In the early 2000s, microarrays were widely used in experiments to monitor the expression profiles of genes in the Arabidopsis genome (Brady et al., 2007; Dash et al., 2012). In addition to genome and transcriptome studies, other omics data such as proteomics and metabolomics also provide crucial insight into the abiotic stress responses (Haak et al., 2017). The Arabidopsis genome sequence and associated omics study are cornerstones of recent significant progresses in abiotic stress research in plants (Hirayama and Shinozaki, 2010).

1.2 MECHANISMS OF STRESS TOLERANCE: GENE EXPRESSION CHANGE

Plants adapt to abiotic stresses through dynamic responses at physiological, biochemical and molecular levels. The development of stress tolerant crops in stress prone areas will be crucial for modern agriculture around the world. In this section, some known molecular mechanisms and pathways related to plant abiotic stress response in the model plant species, *Arabidopsis thaliana,* are reviewed.

In *Arabidopsis thaliana*, during the vegetative growth phase, ABA-regulated gene expression is observed under osmotic stress conditions. During osmotic stress, SnRK2-AREB/ABF pathway plays a key role in the gene expression of ABA response element (ABRE)-dependent pathway (Fujita et al., 2011).

The overproduction of reactive oxygen species (ROS) in plants is detected in various abiotic stresses. These ROS are highly reactive, toxic and can cause damage to major biomolecules including proteins, lipids, carbohydrates and DNA, hence resulting in oxidative stress (Gill and Tuteja, 2010). ROS includes the free radicals (O_2^{*-}), superoxide radicals (OH^*), hydroxyl radical (HO_2^*), per hydroxyl radical (RO^*), non-radical forms such as H_2O_2 (hydrogen peroxide) and singlet oxygen. In normal steady state, there are several anti-oxidative defense mechanisms that scavenge the ROS molecules (Foyer, 2005). Plants can produce antioxidants such as Super Oxide Dismutase (SOD), Catalase (CAT), and glutathione reductase, which act as defense machineries to protect cells from damages caused by oxidative stress. Under stress conditions, the equilibrium between the production and scavenging of ROS radicals is disturbed or changed by different abiotic stresses that lead to sudden increase in the level of ROS that damages cellular structures (Gill and Tuteja, 2010). Plant stress tolerance can be improved by increasing the levels of these antioxidants. In addition to the toxicity effects, ROS are signaling molecules in plant development and abiotic stress responses (Apel and Hirt, 2004). Antioxidants responsive genes and ROS-signaling genes are likely to be differentially expressed during different stresses. In earlier studies, the role of ROS in homeostasis regulation in crop plants has been performed, particularly, the genes that affect abiotic stress resistance in crops have been characterized (You and Chan, 2015).

Calcium plays a role as signaling molecule in a number of abiotic stress signaling pathways, for example, calcium chloride made more effective recovery of cold-hardened rice seedlings

for germination than normal cold-hardened seedlings. It has been found that the calcium concentration is elevated in cold conditions. Such increase occurs as a result of influx of calcium from outside the cell, as well as the release of Ca2+ from intracellular stores. This increased level of calcium induces a signal and transduction via calmodulin, calcium-dependent protein kinases, and other Ca^{2+}-controlled proteins. A wide array of downstream responses come into play and the protection of the plant by the adjustment to the new environmental conditions is established (Knight, 1999).

Drought and high salinity are considered to be the major factors that hamper the growth and productivity of crops. NAC transcription factors, such as ERD1 and RD26, have been shown to function in plant development and biotic and abiotic stress. In Arabidopsis and rice, more than 100 NAC genes have been identified. Overexpression of SNAC in both rice and Arabidopsis have shown increased tolerance in drought (Nakashima et al., 2012).

Heat stress transcription factors (HSFs) are in the central of the stress responsive system in plants. The diversification, functional interactions of HSFs, integration into signaling complex, responded networks have been identified and documented (Palaisa et al., 2017).

Previous studies reviewed in this chapter showed that the changes in gene expression play key roles in response to all abiotic stresses. For example, one essential mechanism of plant drought response is through changes in gene expression (Shinozaki and Yamaguchi-Shinozaki, 2007). In Arabidopsis thaliana, it has been found that gene expression changes alter plant metabolic profiles and the result is increased drought tolerance through drought avoidance (Bechtold et al., 2016). Other abiotic stresses also cause changes in gene expression. For example, high sodium chloride (NaCl) causes rapid, dynamic changes in gene expression through the abscisic acid (ABA) pathway (Zhu, 2002; Fujita et al., 2011). Global

expression profiling was conducted in Arabidopsis plants in order to identify genes of potential importance to cold, salt, and drought tolerance and hundreds of transcripts were found to be differentially expressed under these stress conditions (Stress et al., 2002).

1.3 RNA-SEQ

High throughput sequencing technologies can generate millions of short sequence reads (50 to 150 basepairs) for any given biological sample. Transcriptomes, epigenomes, and genome sequencing can all be performed using high throughput sequencing technologies (Oshlack et al., 2010). RNA sequencing (RNA-Seq) has low noise and high reproducibility that possess several advantages over microarrays as a platform of gene expression. RNA-Seq provides better detection and quantification of alternative splicing isoforms than microarray platforms. Many novel isoforms have been identified by RNA-Seq in plants and other species (Morin et al., 2008; Kim and Salzberg, 2011). Long non-coding RNAs can also be discovered through RNA-Seq (Morin et al., 2008; Trapnell et al., 2009; Fujita et al., 2011; Kim and Salzberg, 2011; Wang et al., 2011). Additionally, previously inaccessible complexities in transcriptome such as allele specific expression, novel promoters, antisense RNA and alternative poly-adenylation sites can be revealed by RNA-Seq (Pan et al., 2008; Wagner et al., 2010).

However, NGS produces complex and large number of data which cannot be directly interpreted by biologists. Large volumes of raw sequencing reads are produced from RNA-Seq experiments. Moreover, reduction of sequencing costs exponentially opens up the door to affordable sequencing which can generate more data in an increasing speed. Such high volume, complex data from RNA-Seq experiments can provide enormous insight into the complicated responses in transcriptome. The number of reads produced from an RNA transcript is a measurement of transcript and gene expression. Computational analyses and

methodologies are critical to interpret the data to gain biological insight from the raw data (Oshlack et al., 2010).

1.4 RNA-SEQ PIPELINE

In all these abiotic stress conditions, it is necessary to identify differentially expressed genes using RNA-Seq experiments. The primary objective for gene expression or transcriptome analysis in many biological studies is to identify differentially expressed genes between samples. Such analysis usually includes a few control samples and a few treated samples. Comparisons that are commonly performed include wild type and mutant strains of the same tissue with or without abiotic stress treatments (Oshlack et al., 2010).

One of the most widely used sequencing technologies is provided by the Illumina Inc. (San Diego, CA) for gene expression profiling. Alternative protocols to produce high throughput sequencing libraries include oxford nano-pore and PacBio sequencing which are promised for much lower cost and longer reads in the future (Wang et al., 2011).

1.5 SOFTWARE TOOLS OF THE RNA-SEQ PIPELINE

1.5.1 READ ALIGNMENT: STAR

The ordered nucleic acids sequences construct genomes linearly, but eukaryotic cells reorganize these sequences in transcriptomes through splicing of non-contiguous exons and create mature transcripts. The detection and characterization of these spliced RNAs have been a critical focus of functional analyses of genomes in both the normal and stressed conditions. High-throughput sequencing experiments produces hundreds of millions of short (36nt) to medium (300nt) length sequence reads and it is very challenging to detect where in the genome these reads were originated from. It is also difficult to share such information across research community. There are several algorithms developed to align these reads along

reference genome. However, mapping error rate is high and speed is low for several available several RNA-Seq aligners (Engström et al., 2013). Spliced Transcripts Alignment to a Reference (STAR) software uses sequential mapping to speed up the search process. It implemented suffix arrays with seed clustering and stitching. The mapping speed and accuracy is higher as compared to other conventional mapping tools. STAR can also map full length RNA sequences with an 80–90% success rate. Due to these advantages, STAR was selected as the mapping software for our pipeline. The STAR software is implemented as a standalone C++ program (Dobin et al., 2013).

1.5.2 READ COUNTING: FeatureCounts

Various genomic analyses require read summarization, which is the process of counting of reads that aligned to genomic regions which are also called genomic features. FeatureCounts is a read summarization program, which is suitable for counting reads of RNA or genomic DNA sequencing experiments with highly efficient chromosome hashing and feature blocking. It works faster and requires less computer memory than most other programs with similar function (Anders et al., 2014). Both single and paired reads can be counted by FeatureCounts with multiple options and appropriate for various applications of sequencing (Liao et al., 2014).

1.5.3 DIFFERENTIAL EXPRESSION: DESeq2

High throughput sequencing technologies provide a large volume of sequence data that needs to be analyzed statistically to find the differentially expressed genes (DEGs) between groups of samples. The comparative transcriptomics analysis is based on the null hypothesis that the logarithmic fold change (LFC) between treatment and control samples is not significantly different from zero. DESeq2 works on shrinkage estimation for dispersions and fold changes

for the differential analysis of count data, which provides improvement of stability and strength of quantitative estimation (Love et al., 2014)

1.5.4 DIFFERENTIAL EXPRESSION: edgeR

A different statistical package used in this study is called edgeR (empirical analysis of DGE in R). Many statistical methods in genomics face major challenges from the high throughput, modern biological data, such as requirement of multiple testing procedures (Robinson et al., 2010). EdgeR uses empirical Bayes to improve inference by sharing information across all observations. EdgeR is a Bioconductor software package. Replicated count-based expression is analyzed by edgeR and the implemented methodology is developed for serial analysis of gene expression (SAGE). It can also be applicable to RNA-Seq, ChIP-Seq and proteomic experiments (Robinson et al., 2010).

1.6 REFERENCES

Anders S, Pyl PT, Huber W (2014) HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics **31**: 166–9

Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annu Rev Plant Biol **55**: 373–399

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

Bechtold U, Penfold CA, Jenkins DJ, Legaie R, Moore JD, Lawson T, Matthews JSA, Vialet-Chabrand SRM, Baxter L, Subramaniam S, et al (2016) Time-series transcriptomics reveals that *AGAMOUS-LIKE22* affects primary metabolism and eevelopmental processes in drought-stressed Arabidopsis. Plant Cell **28**: 345–366

Brady SM, Orlando DA, Lee J-Y, Wang JY, Koch J, Dinneny JR, Mace D, Ohler U, Benfey PN (2007) A high-resolution root spatiotemporal map reveals dominant expression patterns. Science **318**: 801–6

Dash S, Van Hemert J, Hong L, Wise RP, Dickerson JA (2012) PLEXdb: gene expression resources for plants and plant pathogens. Nucleic Acids Res **40**: D1194-201

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**: 15–21

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, et al (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods **10**: 1185–91

Flowers TJ, Garcia A, Koyama M, Yeo AR (1997) Breeding for salt tolerance in crop plants — the role of molecular biology. Acta Physiol Plant **19**: 427–433

Foyer CH (2005) Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses. Plant Cell Online **17**: 1866–1875

Fujita Y, Fujita M, Shinozaki K, Yamaguchi-Shinozaki K (2011) ABA-mediated transcriptional regulation in response to osmotic stress in plants. J Plant Res **124**: 509–525

Gill SS, Tuteja N (2010) Reactive oxygen species and antioxidant machinery in abiotic stress tolerance in crop plants. Plant Physiol Biochem **48**: 909–930

Haak DC, Fukao T, Grene R, Hua Z, Ivanov R, Perrella G, Li S (2017) Multilevel regulation of abiotic stress responses in plants. Front Plant Sci. doi: 10.3389/fpls.2017.01564

Hirayama T, Shinozaki K (2010) Research on plant abiotic stress responses in the post-genome era: past, present and future. Plant J **61**: 1041–1052

Kikuchi A, Huynh HD, Endo T, Watanabe K (2015) Review of recent transgenic studies on abiotic stress tolerance and future molecular breeding in potato. Breed Sci **65**: 85–102

Kim D, Salzberg SL (2011) TopHat-Fusion : an algorithm for discovery of novel fusion transcripts.

Knight H (1999) Calcium signaling during abiotic stress in plants. Int Rev Cytol **195**: 269–324

Liao Y, Smyth GK, Shi W (2014) FeatureCounts: An efficient general purpose program for

assigning sequence reads to genomic features. Bioinformatics **30**: 923–930

Lobell DB, Field CB (2007) Global scale climate–crop yield relationships and the impacts of recent warming. Environ Res Lett **2**: 14002

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 1–21

Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh TJ, McDonald H, Varhol R, Jones SJM, Marra MA (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques **45**: 81–94

Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2012) NAC transcription factors in plant abiotic stress responses. Biochim Biophys Acta - Gene Regul Mech **1819**: 97–103

Oh S-J (2005) Arabidopsis CBF3/DREB1A and ABF3 in Transgenic Rice Increased Tolerance to Abiotic Stress without Stunting Growth. PLANT Physiol **138**: 341–351

Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential expression results. Genome Biol **11**: 220

Palaisa KA, Morgante M, Williams M, Rafalski A, Palaisa KA, Morgante M, Williams M, Rafalskiab A (2017) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci published by : American Society of Plant Biologists (ASPB) Linked references are available on JSTOR for this article : Contrasting Effects of. **15**: 1795–1806

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative

splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet **40**: 1413–1415

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–40

Roy S (2016) Function of MYB domain transcription factors in abiotic stress and epigenetic control of stress response in plant genome. Plant Signal Behav **11**: e1117723

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

Stress C, Kreps JA, Wu Y, Chang H, Zhu T, Wang X, Harper JF, Mesa T, Row M, Diego S, et al (2002) Transcriptome changes for Arabidopsis in response to. Society **130**: 2129–2141

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

Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T, Blanchette M (2010) Computational analysis of whole-genome differential allelic expression data in human. PLoS Comput Biol **6**: 24

Wang H, Wang H, Shao H, Tang X (2016) Recent advances in utilizing transcription factors to improve plant abiotic stress tolerance by transgenic technology. Front Plant Sci. doi: 10.3389/fpls.2016.00067

Wang L, Si Y, Dedow LK, Shao Y, Liu P, Brutnell TP (2011) A low-cost library construction protocol and data analysis pipeline for illumina-based strand-specific

multiplex RNA-seq. PLoS One. doi: 10.1371/journal.pone.0026426

Yoshida T, Fujita Y, Sayama H, Kidokoro S, Maruyama K, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2010) AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. Plant J **61**: 672–85

You J, Chan Z (2015) ROS regulation during abiotic stress responses in crop plants. Front Plant Sci **6**: 1–15

Zhu JK (2002) Salt and drought stress signal transduction in plants. Annu Rev Plant Biol **53**: 247–273

CHAPTER 2

RNA-SEQ ANALYSIS OF PUBLISHED RNA-SEQ DATA FOR *ARABIDOPSIS THALIANA* UNDER DIFFERENT ABIOTIC STRESSES

2.1 INTRODUCTION[2]

A detailed review of abiotic stress responses in plants and RNA-Seq technology were provided in chapter one. A brief summary is provided here for the completeness of this chapter.

Abiotic stresses, including drought, high salinity, heat, and flooding, are major constraints to agricultural productivity worldwide. For example, drought severely reduces crop yield, because the physiology, growth, development, and productivity of crops are largely affected by drought (Lobell and Field, 2007). One essential mechanism of plant drought response is through changes in gene expression and it is crucial to understand this mechanism in order to improve plant performance under drought conditions (Shinozaki and Yamaguchi-Shinozaki, 2007). In *Arabidopsis thaliana*, it has been found that gene expression changes alter plant metabolic profiles and the result is increased drought tolerance through drought avoidance (Bechtold et al., 2016). Other abiotic stresses also cause changes in gene expression. For example, high salinity is another type of abiotic stress, which causes damage to plants through ionic and osmotic stress (Flowers et al., 1997). High sodium chloride (NaCl) causes rapid, dynamic changes in gene expression through the abscisic acid (ABA) pathway (Zhu, 2002; Fujita et al., 2011)

[2] Authors: Shamima Akter, Jiyoung Lee, Hazem Sharaf, Song Qi and Dr. Song Li. Author contributions: Shamima Akter developed the pipeline and prepared the first draft of this chapter. Shamima Akter, Jiyoung Lee, Hazem Sharaf, Song Qi and Dr. Song Li contributed to the development of the pipeline. Shamima Akter and Dr. Song Li edited this chapter.

RNA-Seq analysis is becoming more popular due to its high reproducibility, low noise and is replacing microarray experiments for transcriptome profiling. It is widely used to identify gene expression changes in response to abiotic and biotic stresses, in different tissue types and developmental stages. RNA-Seq is also capable of detecting novel isoforms and long intergenic noncoding RNAs (lincRNAs) (Ching et al., 2014), which are potentially new regulators of abiotic stress responses.

RNA-Seq experiments can generate a lot of raw data. To transfer such huge amount of data into biological information, it requires high-performance computing environments, which are generally powered by Linux operating system. A data processing pipeline usually consists of multiple steps, and each step requires specialized software. Due to the complexity of data processing for RNA-Seq, it is important to develop a flexible and reproducible pipeline to facilitate the data analysis process and to obtain biologically meaningful information from the raw data files.

2.2 OBJECTIVES

The first objective of the study is to develop a RNA-Seq analysis pipeline and to use this pipeline to re-analyze published RNASeq data. The second objective is to perform meta-analysis of published RNA-Seq data for *Arabidopsis thaliana* under different abiotic stresses and to identify stress responsive genes under these abiotic stresses.

The main goals of this study include 1) To identify common stress responsive genes across multiple stresses; 2) To identify unique responsive genes for each stress; 3) To find the similarity of stress responsive genes between different stresses and 4) To analyze the functions of these common and unique stress responsive genes. Therefore, the overall objective of this study is to identify common and unique stress regulated genes and to

understand the functional roles of these genes. A potential future application of this study is to apply the knowledge of stress responsive genes in the breeding of stress tolerance crops. To achieve this long-term goal, future research will be performed to compare Arabidopsis genes to other species to find conserved mechanisms for stress responses. In this study, we developed a fully automated pipeline that produces standardized outputs of all publicly available RNA-Seq data in abiotic stress-related experiments of Arabidopsis before July 2017. Among all these RNA-Seq datasets, we have selected data for eight different abiotic stresses to identify differentially expressed genes under these conditions.

2.3 METHODS

2.3.1 PROCESS OF RNA-Seq ANALYSIS

The first step of RNA-Seq analysis is experimental design. A typical experiment is designed to have two types of conditions including control and treatment for a given type of stress. Control condition includes the plants growing under normal conditions whereas the treatment condition includes plants treated with various abiotic stresses including heat, cold, wounding, drought or combination of two or multiple stresses. For both control and treatment conditions, biological replicates are typically included because replications are required for proper statistical analysis. After mRNA extraction and cDNA synthesis, libraries from samples under different conditions are prepared for sequencing. In the second step, all libraries are sequenced through next generation sequencing platforms which generated sequencing reads in FASTQ files. In the third step, sequenced reads are analyzed through the RNA-Seq pipeline using high performance computing system. In the end, differentially up- or down-regulated genes are obtained using statistical analysis of read counts. Up regulated genes are genes that expressed higher in treated plants as compared to control plants. The down-regulated genes are genes that expressed lower than treated plants as compared to

control plants.

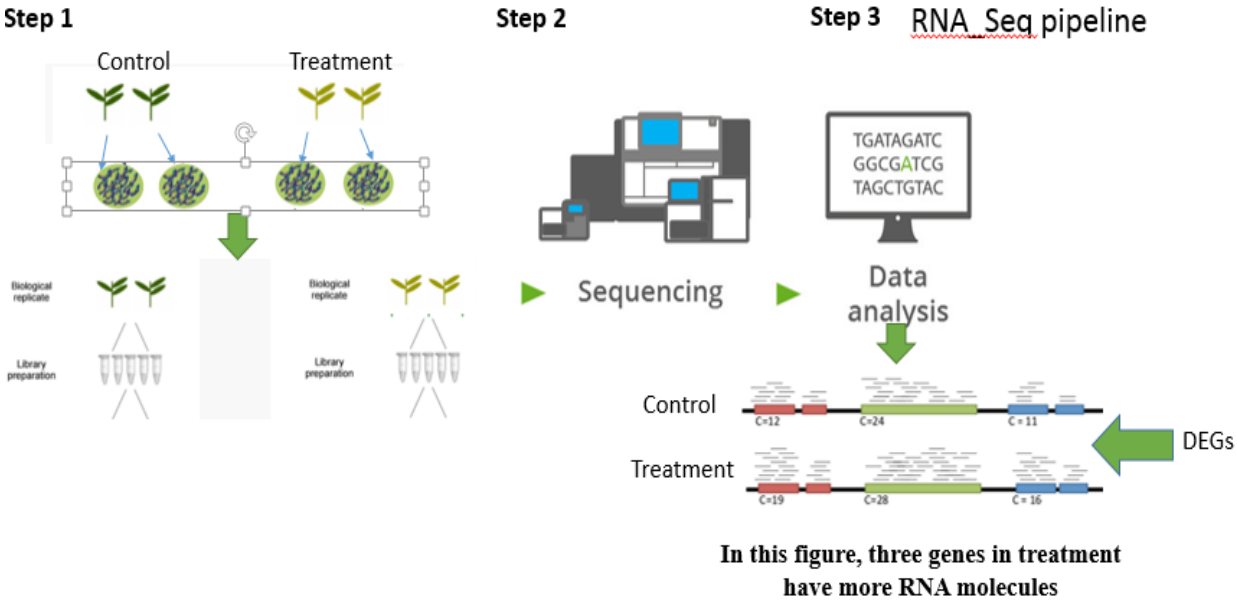

Figure 2.1 Process of RNA-Seq analysis. Three steps in RNA-Seq analysis: 1) Experimental design 2) Sequencing 3) Data analysis in high performance computing system.

2.3.2 PIPELINE OVERVIEW

The RNA-Seq pipeline uses sequencing libraries (raw reads) in fastq format with biological replicates for all treatments. The reads are subsequently analyzed by several software tools: 1) ASCP, 2) Fastq-dump, 3) Cutadapt, 4) STAR, 5) FeatureCounts, 6) DESeq2 and edgeR (see section 2.2.5 for details).

The pipeline develop in this project includes six steps: data download, quality control, read alignment, read counting, expression quantification, and differential expression. In this study, the latest genome and annotation files of *Arabidopsis thaliana* were used and obtained from Araport 11 (Cheng et al., 2017). Data download link is https://www.araport.org/downloads/Araport11_Release_201606/annotation

Differential expression of genes is identified by comparing samples under each abiotic stress to corresponding control samples. Gene expressions are summarized as mean FPKM (Fragments Per Kilobase per Million reads) for each biological condition. In this project, a systematic and comprehensive query were performed in July 2017 and followed by manual inspection to identify all abiotic stress related experiments published in Arabidopsis. In total, 652 published RNA-Seq datasets of Arabidopsis thaliana from GEO database were identified and analyzed.

2.3.3 FLOW CHART OF THE PIPELINE

In first step, raw data were downloaded from Gene Expression Omnibus database as SRA files using **ASCP** software in Advanced Research Computing (ARC) server at Virginia Tech and all output data are as saved as "SRR accession id".sra. **ASCP** is a software developed by IBM, which support very fast, BIGDATA download and upload in cloud-based computing platforms. After downloading the SRA files, **fastq-dump** were used to convert SRA files into fastq data under the file names such as "SRR accession id".fastq. The third step is the quality control step to remove adapter sequences from the raw data by cutadapt (Martin, 2011), which is a software tool that removes adapter sequences from the raw sequencing reads. The mapping of the preprocessed data was done on the reference genome of Arabidopsis through **STAR** and output file is in the bam format (Li et al., 2009), which provides alignment information for each read. The fifth step is to count reads of aligned reads and this step was performed by **featureCounts** to generate readcount.txt files. This is an important step in quantifying the relative expression levels of each gene in a sample. In the sixth step, the merging of read counts were completed in **R** programming language and results are merged for each experiment. In the last step, finding of the differentially expressed genes (**DEGs**) was performed using **DESeq2** and then FPKM was calculated by **edgeR.**

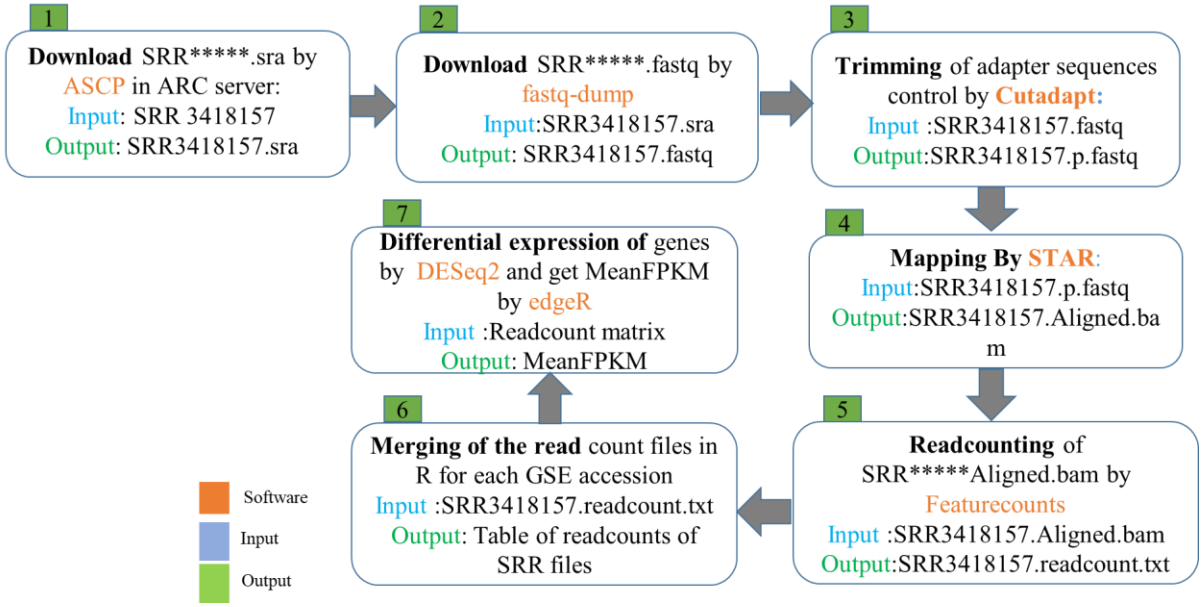

Figure 2.2 Flowchart of the RNA-Seq pipeline. There are seven steps in this pipeline. In each step, software names were denoted by orange text. In every step one input file is needed which is represented by blue text whereas the output file is denoted by green text. The output file of each step is the input for the next step.

2.3.4 SCRIPTS OF THE PIPELINE

The scripts for this pipeline were developed using bash scripting language. All scripts are included in appendix 1, 2, 3, and 4. All scripts and softwares used in this pipeline will be uploaded to Github (https://github.com/orgs/LiLabAtVT/) for the public. The pipeline includes the following scripts: 1) Data download scripts by ASCP and data extraction by fastq-dump; 2) Genome index script, which generates genomic index for reading mapping; 3) Read mapping script for mapping RNA-Seq reads to genome by STAR; 4) Read counting script by featureCounts; and 5) An R script for merging read counts, getting differential expression of genes, and obtaining FPKM.

2.3.5 SOFTWARES OF THE PIPELINE

2.3.5.1 ASCP FROM ASPERA

A standard network protocol for exchanging files across the Internet is the File Transfer Protocol (FTP). However, FTP protocol has slower transfer speed as compare to Aspera, a file transfer protocol that is hundreds of times faster than FTP. It provides greater user control that enables them to set individual transfer rates and bandwidth. Using ASCP, 652 samples have been downloaded from SRA database (Alnasir and Shanahan, 2015).

Short Read Archive (SRA) database contains all publicly available short read data https://www.ncbi.nlm.nih.gov/sra/docs/. Each sample is represented by an SRA file with an unique SRR id. For example, SRR3418157.sra is an SRA file and the SRR id is SRR3418157. This sample is an RNA-Seq data of Arabidopsis leaves for an induced ABA-responsive transcription factor. The metadata of this sample can be obtained by searching SRA database at https://www.ncbi.nlm.nih.gov/sra. One biological experiment usually includes several SRA files representing multiple biological replicates and experimental conditions.

2.3.5.2 FASTQ-DUMP

This tool converts SRA data into fastq format. FASTQ is a format widely used by sequencing data. Each record of a fastq file contains four rows. The first row starts with @ and follows by a sequence identifier that is unique to each read. The second row is read sequence represented by the standard alphabet of DNA sequences (A,T,C,G for the standard nucleotides and other characters that represent nucleotide combinations). The third row starts with '+' and follows by, sometimes, the sequence identifier or simply an empty line. The fourth row is quality values for the sequence. Using this tool, we have extracted 652 fastq

files from corresponding SRA files (both paired and single end). For each single end experiment, we obtained one file in the format of SRRXXXXXX.fastq, with X representing a single integer. For paired end reads, we obtained two files in the format of SRRXXXXXX_1.fastq and SRRXXXXXX_2.fastq.

2.3.5.3 CUTADAPT

This is a quality control tool to remove adapter sequences from the end of sequence reads.

2.3.5.4 STAR

Spliced Transcripts Alignment to a Reference (STAR) is an ultrafast universal RNA-Seq aligner. It is a new algorithm for aligning high-throughput long and short RNA-Seq data to a reference genome. STAR is capable of running parallel threads on a multicore system and provides better alignment precision and sensitivity than other RNA-Seq aligners for both experimental and simulated data (Dobin et al., 2013). STAR maps all the reads on the genome and produces SAM or BAM files and summary statistics files. SAM format stands for Sequence Alignment Map and it is a standardized format for short read alignment data (Li et al., 2009). BAM format is the binary file for SAM files, and BAM files are usually of smaller file sizes as compared to SAM files with the same content. For each read, the SAM file provides information regarding where the read is mapped to and on which chromosome among other information such as read mapping quality. The location and chromosome name of a read is the information we are interested in because we can use such information to count how many reads are from each gene. In our pipeline, we only work with uniquely aligned reads to minimize ambiguity.

2.4.5.5 FeatureCounts

FeatureCounts is a program that counts mapped reads for genomic features such as genes, exons, promoter and chromosomal locations and works highly efficiently (Liao, Smyth, & Shi, 2014). In our experience, it can count 0.5 million reads per second and it took 30-45 seconds to count reads for one RNA-Seq mapped data file such as SRR3418157, which is one RNA-Seq sample generated in a DEX-induced GFP construct control experiment from the data collection in this study. FeatureCount was used to count RNA-Seq reads mapped to exon regions of each gene. In this analysis, the focus is on gene level expression analysis and not to calculate isoform expression levels, because unlike gene expression, there are many isoform expression quantification methods and there is no single software that performs best in this task.

2.3.5.6 DESeq2 AND edgeR

These are R Bioconductor packages. DESeq2 (Love, Huber, & Anders, 2014) analysis identifies differentially expressed genes between conditions. The input data are read counts generated by featureCounts. edgeR can calculate FPKM for each gene (Robinson et al., 2010). The negative binomial distribution is the basis for differential gene expression analysis for both DESeq2 and edgeR. All analysis in this study were performed using Advanced Research Computing (ARC) servers.

2.3.6 DATA SETS

Data Item	Count
Experiments	43
Time series	3
Mutant	29
Runs	652
Samples	554
Conditions	260
Stress	8
Ecotypes	18
Tissue Types	19

Table 2.1 This table summarizes the 43 experiments of published data for *Arabidopsis thaliana* have been collected from GEO database. In all these data, there are 3 time series experiments, 652 SRRs with 260 conditions, 8 stress types and 19 tissue types.

Name of the experiment as Gene Omnibus series	Name of the stress
GSE59699 (Pedmale et al., 2017)	Blue Light
GSE60865 (Suzuki et al., 2015)	High Light
GSE63406 (Schlaen et al., 2015)	Cold
GSE69077 (Rawat et al., 2015)	Heat
GSE67332 (Gehan et al., 2015)	Cold
GSE69510 (Sagor et al., 2016)	Salt
GSE72806 (Suzuki et al., 2016)	Salt and Heat
GSE81202 (Kohnen et al., 2016)	Red Light
GSE80565 (Song et al., 2016)	ABA(drought)
PRJNA324514	Wounding

Table 2.2 Eight stresses: blue light, high light, red light, cold, heat, salt, ABA and wounding from 10 experiments were selected. These experiments were selected because they have properly designed biological replicates. One experiment (GSE72806) has combination of salt and heat stresses. One experiment has three combined stresses including heat, cold and wounding. Two separated experiments (GSE63406 and GSE67332) tested the cold stress.

2.4 RESULTS

652 published data sets of *Arabidopsis thaliana* with eight different stress have been analyzed in this study and generated 644 read count files. Among 652 published samples, eight

samples do not contain any reads that can map to the genome, therefore 644 read count files have been generated. Summary for all data collection is presented in Figure 2.3.

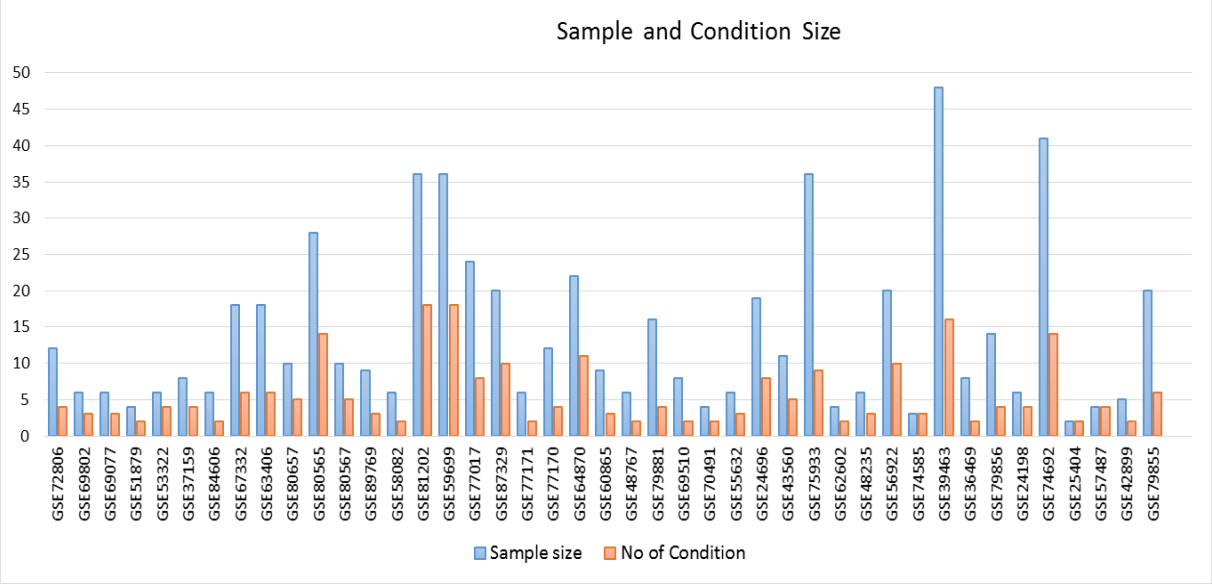

Figure 2.3 Summary of 43 experiments that have been analyzed through the pipeline.

All experiments are presented here as Gene Expression Omnibus Series (GSE) accession ID. Among all of these experiments, 11 experiments have large number of samples and conditions, for example, GSE39463 has 48 samples and 16 conditions, GSE74692 has 42 samples and 14 conditions. All the other experiments have less than 10 number of samples and conditions.

Total number of read mapped and average alignment rate are shown in Table 2.3. There are 16 billion reads from all experiments with an average 25 million reads for each sample that have been analyzed. On average, more than 91% uniquely mapped reads were obtained. There are 20 samples that contain mapped reads that are fewer than 50.0% of total reads. These samples are not included in downstream analysis because the low mapping rate may indicate poor mapping library quality or sample contamination. In Table 2.4, there are 644 samples that were retrieved to for counting reads. Through featureCounts, 15.4 billion reads

and on average 23.0 million reads per sample were assigned for features (exons). Some reads are not assigned to any known features. These reads could have mapped to intron regions or intergenic region, which is commonly found in RNA-Seq experiments.

No. of total input reads for all sample	No. of average input reads per sample	Average of uniquely mapped reads per sample	No. of samples with > 50% uniquely mapped reads
16 billion	25 million	91.88%	20

Table 2.3 Summary of read mapping results from STAR software. There were 652 samples used in this analysis.

Description	Number of samples	Total number of read counts for all samples	Avg. read counts per sample
Assigned	644	15.4 billion	23.0 million
Unassigned	644	0.175 billion	0.272 million

Table 2.4 Summary of reads counted by FeatureCounts software. There were 644 samples used in this analysis.

Among all samples, 6 experiments (GSE series) were selected for the analysis of differentially expressed genes. These samples were selected because these experiments include large number of samples and conditions and proper biological replicates were found in these experiments. After DESeq2 analysis, significant DEGs were identified as genes with adjusted p-value < 0.01. A number of up-regulated genes and down-regulated genes were

obtained from each of these experiments. One experiment, GSE 80567, which is the RNA-Seq of abscisic acid-responsive transcription factors, DIG, has the highest number of up- and down-regulated DEGs. Summary of the DESeq2 analysis is presented in Table 2.5.

Name of experiments	No. of conditions	No. of samples	No. of genes with non-zero total read counts	No. of genes with p value adjusted to <0.01						
				log2 fold change > 0 (up)		log2 fold change < 0 (down)		Low count		No Change
GSE56922	10	20	27,899	3308	12%	2798	10%	6907	25%	53%
GSE59699	18	36	30,014	5592	19%	5505	18%	6893	23%	40%
GSE64870	11	41	30661	4474	15%	3783	12%	9402	31%	42%
GSE74692	14	10	29,213	2362	8%	1978	7%	7996	27%	58%
GSE80565	14	28	28860	1107	4%	1588	6%	7712	27%	64%
GSE80567	5	22	26941	6972	26%	7671	28%	4094	15%	31%

Table 2.5 Summary from the DESeq2 run for six selcted GSE experiments.

Ten abiotic stress experiments were selected to run RNA-Seq analysis to find differentially expressed genes. These data are from ABA, salt, heat, cold, blue light, high light, red light, and wounding experiments. These datasets were selected from 43 experimental data, processed and analyzed through the pipeline, and identified lists of differentially expressed genes for each stress experiment. Genes passed the adjusted p value filter are further categorized based on their log2 fold change into up- and down-regulated genes. All these lists of up- and down-regulated genes were sorted using an R script and these lists are compared to identify commonly up- or down-regulated genes across all stresses. Firstly, any pair of

stresses were compared to identify common genes between any two stresses. Secondly, genes were compared across all stresses to identify genes that are induced or repressed in three or more types of stresses. Finally, the gene lists that are only expressed in one particular stress condition were also obtained. After getting the common and unique genes of each stress, we calculated the number of common and unique stress responsive genes, performed gene functional analysis using Gene Ontology (Gene Ontology Consortium, 2015) and characterized the similarity between stress regulated gene lists using Jaccard index.

To understand how many genes are regulated by individual stress or a combination of stresses, a bar plot (Figure 2.4) was created for common up-regulated genes and down-regulated genes. From this plot, there are 5729 up-regulated and 5062 down-regulated genes in response to a single stress type. These numbers are the sum of number of genes that are only found to be responsive to individual stress type. For example, there are 1843 genes induced by ABA. Among these genes, 290 are only induced by ABA, but not by other streses. These 290 genes are included in the 5729 up-regulated genes that are only induced in one stress type. Similarly, genes that are only induced by heat, cold and other stresses are also included in the list of 5729 genes. The 2nd set of bars shows number of genes induced or repressed by two stresses (4514 and 3723 genes respectively). Using ABA induced genes as an example, among the 1843 ABA induced genes, there are 319 genes induced by ABA and one other stress type. These 319 genes include genes that are induced by ABA and cold, ABA and heat etc. These 319 genes do not include genes induced by more than two stress types. In Figure 2.4, the number of these genes are decreasing with increasing number of stresses, which means the number of DEGs between two stresses are fewer than that of only one stress. There are only 18 up-regulated or induced genes across all stresses, and 12 down-regulated or repressed genes in all stresses. This means there are 18 genes that are up-regulated in all eight different stresses and 12 genes are down-regulated in all eight stresses.

This pattern is understandable because different stresses are likely to induce some stress specific genes as well as some genes, such as oxidative stress responsive genes which are commonly used to reduce stress induced damages to plants. When comparing two different stresses, such as drought and salt stresses, one would expect a large number of commonly induced genes because these stresses affect similar cellular processes. However, when the number of stresses increases, the commonly induced genes will be decreasing because different stresses induced more stress specific genes than common genes.

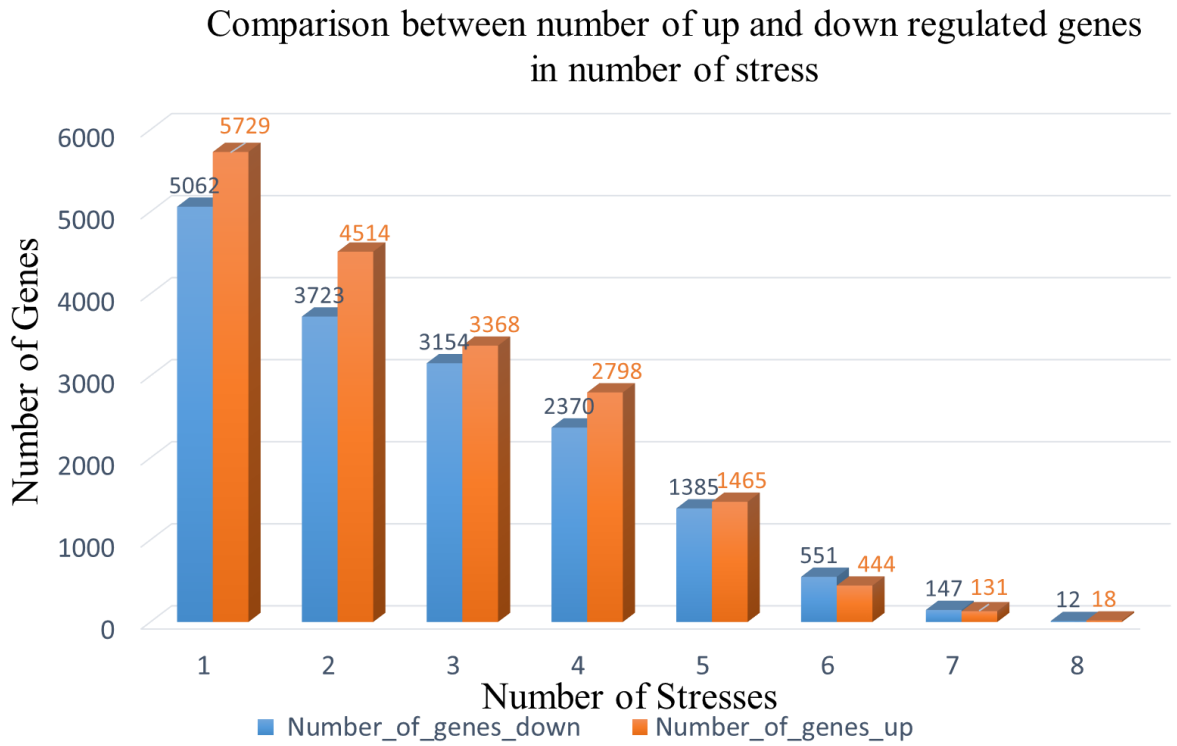

Figure 2.4 Comparison between number of up and down regulated genes in number of stress. X axis represents number of stresses and y axis represents number of genes. Blue colored bar denotes down-regulated and orange colored bar denotes up-regulated genes. This bar plot showed decreasing pattern of number of genes with increased number of stresses.

These commonly induced or repressed genes for all stresses were selected and performed functional analysis. Several abscisic acid responsive genes were induced (Table 2.6).

Notably, ERD10 is a dehydrin family protein which is known to be involved in ABA responses. ABI1 is phosphatase 2C family protein, which is one of the most well characterized ABA responsive genes. RD26 is a transcriptional regulator protein. ABF3 is an abscisic acid responsive elements binding factor 3. These genes with known stress related functions are highlighted in this Table 2.6.

The genes that are commonly repressed among all stresses were identified (Table 2.7). These genes include CAB, a chlorophyll binding protein; AT4G12980 which is auxin responsive protein; DIN10, which is raffinose synthase family protein; RGF9 is a root meristem growth factor protein; Integrase type DNA binding protein; and CYS-HIS rich thioredoxin. Genes with known roles in stress responses are highlighted in the Table 2.7. These genes are known to be down-regulated in response to ABA. In summary, many genes that were known to be involved in plant stress responses were identified by the RNA-Seq pipeline. Several genes that have not been functionally characterized were also identified by our analysis.

Gene name	Known name	Function
AT1G02660	AT1G02660	alpha/beta-Hydrolases superfamily protein
AT1G13990	AT1G13990	plant/protein
AT1G20450	ERD10	Dehydrin family protein
AT1G32900	GBSS1	UDP-Glycosyltransferase superfamily protein
AT1G60590	AT1G60590	Pectin lyase-like superfamily protein
AT1G73480	AT1G73480	alpha/beta-Hydrolases superfamily protein
AT1G78850	AT1G78850	D-mannose binding lectin protein with Apple-like carbohydrate-binding domain-containing protein
AT2G39800	P5CS1	delta1-pyrroline-5-carboxylate synthase 1
AT2G41870	AT2G41870	Remorin family protein
AT3G55500	EXPA16	expansin A16
AT4G04020	FIB	fibrillin
AT4G14270	AT4G14270	polyadenylate-binding protein interacting protein
AT4G26080	ABI1	Protein phosphatase 2C family protein
AT4G27410	RD26	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein
AT4G29190	OZF2	Zinc finger C-x8-C-x5-C-x3-H type family protein
AT4G34000	ABF3	abscisic acid responsive elements-binding factor 3
AT5G01600	FER1	ferretin 1
AT5G01820	SR1	serine/threonine protein kinase 1

Table 2.6 Functional annotation of common stress responsive genes obtained using Thalemine tool in Araport https://apps.araport.org/thalemine/.

Gene name	Known name	Function
AT1G29910	CAB3	chlorophyll A/B binding protein 3
AT1G58602	AT1G58602	LRR and NB-ARC domains-containing disease resistance protein
AT1G66970	SVL2	SHV3-like 2
AT1G76160	sks5	SKU5 similar 5
AT3G14310	PME3	pectin methylesterase 3
AT4G12980	AT4G12980	Auxin-responsive family protein
AT5G20250	DIN10	Raffinose synthase family protein
AT5G54710	AT5G54710	Ankyrin repeat family protein
AT5G61440	ACHT5	atypical CYS HIS rich thioredoxin 5
AT5G61590	AT5G61590	Integrase-type DNA-binding superfamily protein
AT5G62360	AT5G62360	Plant invertase/pectin methylesterase inhibitor superfamily protein
AT5G64770	RGF9	root meristem growth factor

Table 2.7 The functional annotation of the common repressed genes using Thalemine tool in Araport https://apps.araport.org/thalemine/. Highlighted genes are known to be involved in plant stress responses and are discussed in main text.

Known Name	Activity
AT1G02660	Lipase activity in lipid metabolism
ERD10	The stress tolerance in cold and drought stress
GBSS1	Glucan biosynthetic process, starch synthase
P5CS1	Delta1-pyrroline-5-carboxylate synthetase activity in in salt stress relating proline accumulation, ROS protection
EXPA16	Cell wall modification involved in multidimensional and unidimensional cell growth, plant-type cell wall loosening
FIB	Structural molecule activity in photo inhibition, response to cold, abscisic acid stimulus, involved in abscisic acid-mediated photo protection
ABI1	Positively acts on the stomatal closure that explains one important stress response
RD26	Transcriptional activator in ABA-mediated dehydration response
ABF3	Protein binding, DNA binding, transcription activator activity and transcription factor activity
SR1	Protein amino acid phosphorylation, Protein serine/threonine kinase activity, protein kinase activity, ATP binding in signal transduction
CAB3	Chlorophyll binding in photosynthesis, light harvesting, photosynthesis
AT4G12980	Auxin-responsive family protein
AT1G58602	ATP binding in defense response, apoptosis
sks5	Oxidoreductase activity, copper ion binding in oxidation reduction
ACHT5	Cell redox homeostasis
AT5G61590	Negatively regulates cuticular wax biosynthesis
RGF9	Maintenance of the root stem cell niche and transit amplifying cell proliferation.

Table 2.8 The detailed functions of the common induced and repressed genes. The genes in the light blue area are the induced genes and the genes in the light orange are repressed genes.

The functional analysis of the common up and down regulated genes was done in the ThaleMine tool in Araport (Cheng et al., 2017). The gene, AT1G02660, encodes protein with lipase activity in lipid

metabolism pathway. Activation of this genes is associated with the reduction of energy production from photosynthesis. GBSS1 has starch synthase activity, which is also an indication of reduced photosynthetic activity. ERD1 (early response dehydrin 1) has activity of stress tolerance in cold and salt tolerance. P5CS1 has function in reactive oxidative species (ROS) protection, which plays protective roles under stress conditions. EXPA16 is a cell wall modification gene that supports the cell wall expansion, suggesting cell wall remodeling is a key process in plant stress tolerance. The activity of fibrin protein is involved in sensing stress stimulus. ABF3 is a transcription factor in ABA signal transduction. ABF3 encodes abscisic acid responsive element binding protein in response to stress and ABA that has been know to be a central regulatory gene in plant stress responses (Yoshida et al., 2010). The SRI protein has serine/threonine kinase activity in ABA signal transduction for protein amino acid phosphorylation, which is a key step in signaling transduction pathways in plant stress responses.

For the commonly repressed genes across all stresses, gene functions were also analyzed. The repressed gene CAB3 encodes protein related to cholorphll biosynthesis, which suggests that abiotic stresses reduces photosynthesis. The glycerol and lipid metabolism processes are repressed, suggesting that changes are involved in reduction of the energy production in these pathways. The oxidation–reduction activity of sks5 and cell redox activity of ACHT5 reduction reflects the inhibition of ROS production. Repression of AT5G61590 and RGF9 suggests the necessity of reducing wax production and reduced cell proliferation.

To determine the similarity between responsive genes in any pair of stresses, the Jaccard similarity index was calculated for comparing genes that were induced under each stress condition. Figure 2.5 shows that at least three groups of stresses can be identified based on hierarchical clustering. From the dendrogram produced by the clustering analysis, ABA, salt and high light induced genes are in one group and they have the least similarity to other genes induced by other stresses. These genes have very low similarities (smaller than 20%) within their own group. Heat and cold stresses induced genes are in one group with highest

similarity. Red light, wounding and blue light induced genes are in another group with moderate similarity.

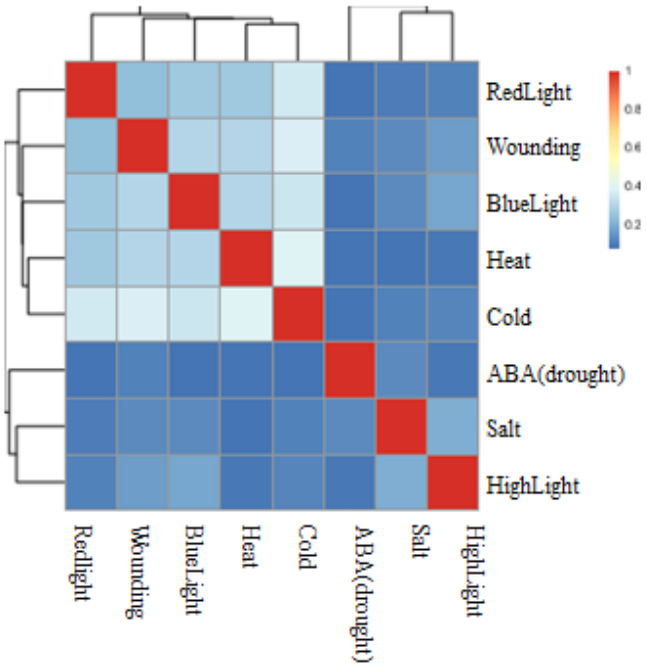

Figure 2.5 The Jaccard similarity index for induced genes for each stress. Jaccard index is defined as Jaccard(A,B) = |A∩B| / |A∪B|. A represents genes in one stress and B represents genes in another stress. Dendrogram is on the left and upper side of the heatmap. Stresses are labeled on both right and lower side of the dendrogram. The upper corner scale denotes the color for similarity. Red color is highest similarity, blue color indicates no similarity and white color denotes moderate similarity. Three groups include group 1: heat, cold; group2: ABA, salt and high light and group 3: blue light, red light and wounding.

The Jaccard similarity analysis for repressed genes between stresses (Figure 2.6) was performed. From the dendrogram and the heatmap, it can be seen that ABA repressed genes has lowest similarity with other stress repressed genes. Three groups were identified in the heatmap. Heat and cold repressed genes are in the same group with high similarity. Red and

blue light repressed genes show moderate similarity. Salt and high light are in the same group with lower similarity than the other two groups.

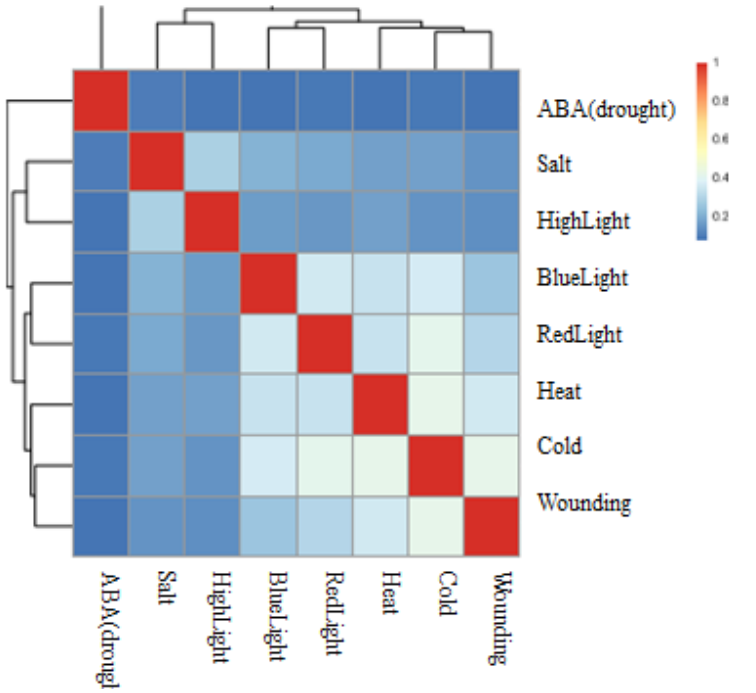

Figure 2.6 Jaccard similarity index for repressed genes for each stress. Dendrogram is in the left and upper side of the histogram. Name of the stresses are in both right and lower side of the dendrogram. The upper corner scale denotes the measurement of similarity: red color for highest similarity, blue for no similarity and white color denotes moderate similarity. Group1: heat, cold; Group2: salt and high light. Group 3: blue light and red light.

The annotations for uniquely induced genes and repressed genes in each stress were analyzed using agriGO, a gene ontology analysis tool. The summary result is presented in Table 2.8.

There are 99 significant gene ontology terms from 16 lists of both induced and repressed genes. However, most of these significant GO terms are generic functions such as

macromolecule binding, which do not provide specific insight into the stress response processes. One exception is that enriched GO terms for ABA up-regulated genes (Table 2.9.1). These gene have function for nutrient reservoir activity and transcription factor activity, suggesting that ABA response is related to nutrient rebalance and the effect of ABA are regulated though transcription regulation.

To identify the commonly induced or repressed gene functional categories across multiple stress conditions, we searched for GO terms that are enriched in multiple stress conditions. This search only resulted in a single GO term (GO:0044425) with gene function for membrane part (Table 2.9.2). This GO term is enriched in repressed genes for 4 stresses (ABA, cold, heat, wounding) and is also enriched in induced genes for two stresses (blue light and red light). Although membrane related regulation are generally considered as part of signaling pathways that mediate fast response to different stresses (Welti et al., 2002; Conde et al., 2011), our result suggests that transcription regulation is also involved in the remodeling of membrane composition under different stress conditions. It is interesting that light induces genes related to membrane activities whereas other stresses repress genes related to membrane activities, implying distinct strategies for plant to couple with stresses.

GO ID	FDR	GO Term
GO:0045735	0.0043	Nutrient reservoir activity
GO:0000988	0.0043	Transcription factor activity, protein binding
GO:0000989	0.0041	Transcription factor activity, transcription factor binding
GO:0044699	0.011	single-organism process

Table 2.9.1 Summary for gene ontology of unique up and down regulated genes in agriGO GO analysis tool.4 GO terms are significant with very low FDR. Genes are in these GO term the genes have Nutrient reservoir, TF, TF binding activity.

Stress type	Induce/repress	GO ID	FDR	GO term	number of genes
ABA_drought	repress	GO:0044425	0.012	membrane part	74
Cold	repress	GO:0044425	0.032	membrane part	266
Heat	repress	GO:0044425	0.01	membrane part	340
Wounding	repress	GO:0044425	0.034	membrane part	83
Blue Light	induce	GO:0044425	0.0003	membrane part	187
Red Light	induce	GO:0044425	1.6E-14	membrane part	342

Table 2.9.2 One single GO term GO:0044425 is significant for the stresses: ABA, cold, heat, wounding with repressed genes and blue light and red light induced genes with function of membrane part. GO annotation was done through AGRiGO go analysis tool.

http://systemsbiology.cau.edu.cn/agriGOv2/

2.5 DISCUSSION

Abiotic stresses are a major concern for agricultural production around the world. Some industrial activities resulted in waste materials that are known as a major contributor to global climate change (Fenger, 2009). Due to these activities, carbon dioxide level rises and causes the increase of temperature resulting in heat and drought stresses (Ramanathan and Feng, 2009). Industrial waste materials could also cause salt stresses (Yadav et al., 2011). These activities can be controlled to reduce emissions but progress has been very slow. Understanding the genetic or molecular mechanisms that are activated under abiotic stresses in crop species is thus a major research goal for plant scientists (Hirayama and Shinozaki, 2010). The availability of the next generation sequencing technologies makes RNA-Seq affordable for studying different crop species (Edwards et al., 2013). The objectives of this study were to identify common and unique stress-responsive genes and to characterize the function of these genes. In the future, the insights of molecular mechanisms of stress tolerance can be potentially applied in the breeding of stress tolerant crops. In this study, RNA-Seq data from *Arabidopsis thaliana* have been analyzed to identify stress responsive genes. A future direction is to expand this research to other species of Brassicaceae family and other plant species and to identify conserved or species specific functional genes related to abiotic stress tolerance.

The developed pipeline is uniformed and fast, because highly efficient computational tools for downloading, mapping and counting reads on to reference genome were used in this pipeline. Although we did not test the download speed of the entire dataset, download individual dataset using ascp is usually more than 10 times faster than downloading using the traditional SFTP software. The mapping step by STAR for all data took 26 hours, and STAR is more than 50 times faster than other existing aligners (Dobin et al., 2013). Counting reads

by featureCounts took 10 hrs, which is also faster than other popular software such as HTseq-count. On average, featureCounts can count reads of one bam file in 5-6 minutes, whereas HTseq-count may take up to 30 minutes to count the same file. This is because HTseq-count requires sorting of the alignment file by read names which is time consuming. The scripts for this pipeline have been designed to be flexible such that the pipeline can be used to analyze data from species other than Arabidopsis. However, once the pipeline is setup to process data, it can process all datasets with uniformed parameters. This is particularly important for the meta-analysis of gene expression data. When these RNA-Seq data were first published, they were processed using different customized pipeline in each publication. However, not all pipelines produce the same results with the same input data due to different software used in the process. Our uniformed processing pipeline eliminates technical bias introduced by different processing pipelines, such that the comparison is not confounded by the pre-processing procedures.

Ten experiments were selected for downstream analysis as these experiments have high mapping rates with proper replicates. After RNA-Seq analysis of 10 experiments of eight abiotic stressed data, common and unique stress genes have been obtained. From the bar plot, it has been found that the number of stress-responsive genes are high in a single stress which suggests that, in any specific stressed condition, a large number of genes are induced or repressed in response to stress to adopt such specific condition. This is also observed that few genes are commonly induced or repressed, which suggests there are some actions that are commonly performed by plants in all stresses. Only a few commonly induced or repressed genes were identified in our analysis, because all the published analyses were only focused on one particular stress. Therefore, the numbers of genes that are overlapping between all these individual treatments are expected to be very small. However, this observation does not rule out the possibility that when plants were treated with multiple stresses simutanously,

many more genes will be affected. In fact, combination of multiple stresses is more common than individual stress in real life. More research in stress combination will likely to reveal novel genes involved in multiple stresses. In our study, ABA-responsive genes have been commonly expressed in all stresses. The common repressed genes lead to reduce energy consumption and metabolism, which is consistent with known function of ABA during seed germination (Garciarrubio et al., 1997; Planes et al., 2015). This strategy is to not making new biomass but maintaining energy by taking some conservative actions to avoid stresses.

The biological implications of stress-regulated genes have been summarized in many review papers, for example, from Roy et al., 2011. According to this article, it is hypothesized that transcription factors such as ABF3, NAC and RD26 induce several genes with functions in the multiple cellular activities such as energy production, cell proliferation, and ROS scavenging. Based on this review article, we have developed a hypothetical model that explains the observed differential expression patterns in our analysis (Figure 2.7).

It is hypothesized that in osmotic, heat and drought stress, the ABA-independent pathway is activated by transcription factors Dehydration Responsive Element Binding protein 1 and 2 (DREB1 and DREB2) (Nakashima et al., 2009). In osmotic or salt stress, the ABA-dependent pathway is also activated by ABREB/ABF that reflects the induction of stress tolerance genes product. However, in our meta-analysis, it is found that across all stresses (heat, cold, highlight, blue light, red light drought, salt, and wounding), ABF3 was induced. Therefore, we can hypothesize that ABA-dependent signal transduction for stress tolerance was engaged in all these stresses. This finding will be useful for further study to test the function of ABF3 in Arabidopsis and in other species such as rice where an ABF3 transgenic line has already been established (Oh, 2005).

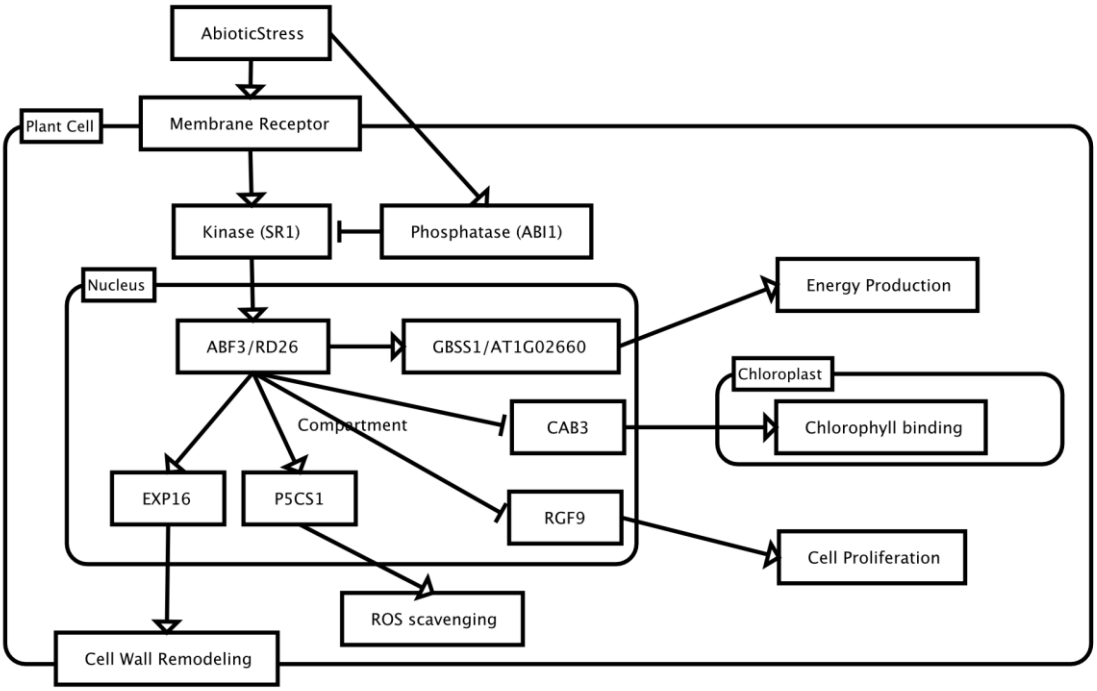

Figure 2.7 A schematic model of abiotic stress response based on genes found in our analysis. Transcription of many genes results in different mechanisms to stress tolerance. In this study of abiotic stress data analysis of *Arabidopsis thaliana*, ABF3/RD26 encodes ABREB that induced genes (GBSS1 and AT1G02660) that positively regulate energy production. P5CS1 was induced by stress and has ROS scavenging activity. EXPA16 has cell expansion or cell wall modification activity. SRI kinases work in ABA signal transduction pathway. Stress also inhibit CAB3, a positive regulation of chloroplast activity and RGF9, a positive regulator of cell proliferation. This model is drawn using BEACON pathway editor (Elmarakeby et al., 2017).

Meta-analysis has been used to study abiotic stress responses. For example, drought, cold, high salinity stresses induced and repressed genes were identified through microarray study

(Seki et al., 2002). This study identified six DREB family transcription factors, two ERF family transcription factors, two NAC family transcription factors, 10 zinc finger family transcription factors, five NAC family transcription factors. In addition to these transcription factors, other genes such as P5CS5, ERD21, RD21, ERD10, RD17, RD29A, RD29B. ERD6, ERD4, RD28, three lipases, two protein ser/Thr protein kinases, two protein kinases, ABI1, two protein phosphatase 2c family proteins, RD20, ERD7, Ferritin, two UDP-glucose glucosyltransferases were up-regulated across three stresses. In this study, the down-regulated genes were ribulose 1,5-bisphosphate carboxylase, 13 chlorophyll a/b-binding proteins, DNA-damage repair protein, pectin esterase, 2-cys peroxiredoxin.

In a recent study (Yoshida et al., 2014), it has been found that ABRE (ABA-responsive element) is cis-acting in the promoter and AREB (ABRE-binding factors) are the transcription factors that plays a pivotal role in ABA-dependent gene expression in osmotic stress. In this study it was also reported that dehydration-responsive element (DRE) is another is cis-acting element in ABA-dependent gene expression and DREB2 (dehydration responsive element binding protein 2) is the transcription factors that bind to these cis-elements (Yoshida et al., 2014). In one study of ABA induced genes, it has been found that glycine-rich protein, no apical meristem protein family (NAC), ERD4, zinc finger protein, glycosyl transferase family protein 20, lipase, ABA insensitive 5(ABAI5), protein phosphatase 2C family protein (Zhang et al., 2005). In an earlier study, it was found that among stress responsive genes, there are four transcriptional regulatory systems, two of them are ABA-independent and two are ABA-dependent (Shinozaki et al., 2003). By genetic and molecular analyses, overlaps between these regulatory systems have been suggested. The ABA-dependent pathways controls the drought-inducible expression of CBF4 suggesting that CBF4 can function that depend on the accumulation of ABA in response to drought slowly (Shinozaki et al., 2003)

Comparing to these earlier studies, our new analysis was performed using RNA-Seq, which can detect all genes in the Arabidopsis genome. Earlier studies using microarrays do not cover all genes in the genome. For example, the widely used ATH1 array only includes 22,000 genes, whereas in the Arabidopsis genome, there are 27,000 genes. Our analysis provides a valuable new finding that across the eight stresses (ABA_drought, highlight, red light, blue light, heat, cold, salt, wounding), there are several novel up-regulated genes were identified such as ERD10, RD26, FIB, EXPA16, and common down regulated genes were raffinose synthase, Sks5, ACTH5 and RGF9. These genes cannot be identified by previous array-based studies is probably because some of these genes were not included in the array platforms used in earlier studies.

It has been difficult to characterize the gene function responsible for stress responses using gene ontology analysis. Only ABA responsive genes have enriched GO categories that are interpretable and are known to be related to stress responses. This result suggests other gene functional annotation is probably needed to better characterize genes in these diverse stress conditions. We found that genes with the function “membrane part” are enrichend in six stress conditions. Although the function of cellular membrane modification has been reported to be related to multiple stress conditions (Haak et al., 2017), there has not been a report to show that changes in membrane activity in involved in all six different stresses. It is interesting that these membrane-related genes are repressed and some are induced. More studies will be required to understand the biological implications of all these genes to gain better mechanical insight into the functional roles of membrane related genes in stress responses.

2.6 CONCLUSIONS

In this study, a RNA-Seq pipeline was developed and was applied to process and analyze a large amount of published RNA-Seq data, including 43 experiments with 652 data set from the model plant species, *Arabidopsis thaliana*. After RNA-Seq analysis of ten selected abiotic stress experiments of *Arabidopsis thaliana*, it was found that 12 differentially expressed genes were repressed and 18 differentially genes were induced across all 8 abiotic stresses.

From the functional analysis of commonly up and down regulated genes, it has been found that these genes are mostly ABA responsive genes following ABA dependent signal transduction. The Jaccard similarity index shows that heat and cold responsive genes are in the same group for both up- and down-regulated genes. High light responsive genes are different from blue light and red light responsive genes. ABA responsive genes have very low similarity with any other stress responsive genes. Four GO terms (GO:0045735, GO:0000988,GO:0000989, and GO:0044699) are significantly enriched in ABA up-regulated genes. These specific GO functions are nutrient reservoir activity, transcription factor, and transcription factor binding activity. One significant GO term, membrane part activity is significant enriched for repressed or induced genes for six stresses.

2.7 FUTURE DIRECTIONS

In the future, stress responsive genes of heat and cold stresses will be analyzed to better understand why changes in temperature from optimal growth temperature can induce and repress similar sets of genes. Additional functional studies of genes that responded to high light will be compared with genes that responded to red and blue lights. Most importantly, we would be interested in understand the roles of ABA responsive genes and why ABA responsive genes are not similar to other stress responsive genes. This is particularly puzzling

because ABA responsive transcription factor such as ABF3 is induced by all stresses. One hypothesis is that ABF3 is the main transcription factor that mediated all stress responses, whereas other transcription factors are specific to different stresses. A potential new avenue to explore is to identify transcription co-regulators using newly developed bioinformatics software (Song et al., 2017). In this study, only gene expression levels were analyzed across multiple stresses. In the future, we will also expand this analysis to splicing isoforms (Aghamirzaie et al., 2016), long non-coding RNAs (Li et al., 2016) and regulatory networks (Redekar et al., 2017) to identify common and uniquely regulated genes across multiple stress conditions.

2.8 REFERENCES

Aghamirzaie D, Collakova E, Li S, Grene R (2016) CoSpliceNet: a framework for co-splicing network inference from transcriptomics data. BMC Genomics **17**: 845

Alnasir J, Shanahan HP (2015) Investigation into the annotation of protocol sequencing steps in the sequence read archive. Gigascience **4**: 23

Anders S, Pyl PT, Huber W (2014) HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics **31**: 166–9

Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annu Rev Plant Biol **55**: 373–399

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

Bechtold U, Penfold CA, Jenkins DJ, Legaie R, Moore JD, Lawson T, Matthews JSA, Vialet-Chabrand SRM, Baxter L, Subramaniam S, et al (2016) Time-series transcriptomics reveals that *AGAMOUS-LIKE22* affects primary metabolism and developmental processes in drought-stressed Arabidopsis. Plant Cell **28**: 345–366

Brady SM, Orlando DA, Lee J-Y, Wang JY, Koch J, Dinneny JR, Mace D, Ohler U, Benfey PN (2007) A high-resolution root spatiotemporal map reveals dominant expression patterns. Science **318**: 801–6

Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD (2017) Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J **89**: 789–804

Ching T, Huang S, Garmire LX (2014) Power analysis and sample size estimation for

RNA-Seq differential expression. 1684–1696

Conde A, Chaves MM, Geros H (2011) Membrane transport, sensing and signaling in plant adaptation to environmental stress. Plant Cell Physiol **52**: 1583–1602

Dash S, Van Hemert J, Hong L, Wise RP, Dickerson JA (2012) PLEXdb: gene expression resources for plants and plant pathogens. Nucleic Acids Res **40**: D1194-201

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**: 15–21

Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet **126**: 1–11

Elmarakeby H, Arefiyan M, Myers E, Li S, Grene R, Heath LS (2017) Beacon editor: capturing signal transduction pathways using the systems biology graphical notation activity flow language. J Comput Biol cmb.2017.0095

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, et al (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods **10**: 1185–91

Fenger J (2009) Air pollution in the last 50 years - from local to global. Atmos Environ **43**: 13–22

Flowers TJ, Garcia A, Koyama M, Yeo AR (1997) Breeding for salt tolerance in crop plants — the role of molecular biology. Acta Physiol Plant **19**: 427–433

Foyer CH (2005) Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses. Plant Cell Online **17**: 1866–1875

Fujita Y, Fujita M, Shinozaki K, Yamaguchi-Shinozaki K (2011) ABA-mediated transcriptional regulation in response to osmotic stress in plants. J Plant Res **124**: 509–525

Garciarrubio A, Legaria JP, Covarrubias AA (1997) Abscisic acid inhibits germination of mature Arabidopsis seeds by limiting the availability of energy and nutrients. Planta **203**: 182–187

Gehan MA, Park S, Gilmour SJ, An C, Lee CM, Thomashow MF (2015) Natural variation in the C-repeat binding factor cold response pathway correlates with local adaptation of Arabidopsis ecotypes. Plant J **84**: 682–693

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

Gill SS, Tuteja N (2010) Reactive oxygen species and antioxidant machinery in abiotic stress tolerance in crop plants. Plant Physiol Biochem **48**: 909–930

Haak DC, Fukao T, Grene R, Hua Z, Ivanov R, Perrella G, Li S (2017) Multilevel regulation of abiotic stress responses in plants. Front Plant Sci. doi: 10.3389/fpls.2017.01564

Hirayama T, Shinozaki K (2010) Research on plant abiotic stress responses in the post-genome era: past, present and future. Plant J **61**: 1041–1052

Kikuchi A, Huynh HD, Endo T, Watanabe K (2015) Review of recent transgenic studies on abiotic stress tolerance and future molecular breeding in potato. Breed Sci **65**: 85–102

Kim D, Salzberg SL (2011) TopHat-Fusion : an algorithm for discovery of novel fusion

transcripts.

Knight H (1999) Calcium signaling during abiotic stress in plants. Int Rev Cytol **195**: 269–324

Kohnen M V., Schmid-Siegert E, Trevisan M, Petrolati LA, Sénéchal F, Müller-Moulé P, Maloof J, Xenarios I, Fankhauser C (2016) Neighbor detection induces organ-specific transcriptomes, revealing patterns underlying hypocotyl-specific growth. Plant Cell **28**: 2889–2904

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup. Bioinformatics **25**: 1–2

Li S, Yamada M, Han X, Ohler U, Benfey PN (2016) High resolution expression map of the Arabidopsis root reveals alternative splicing and lincRNA regulation. Dev Cell **in press**: 508–522

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

Lobell DB, Field CB (2007) Global scale climate–crop yield relationships and the impacts of recent warming. Environ Res Lett **2**: 14002

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 1–21

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: 10

Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh TJ, McDonald H,

Varhol R, Jones SJM, Marra MA (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques **45**: 81–94

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

Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2012) NAC transcription factors in plant abiotic stress responses. Biochim Biophys Acta - Gene Regul Mech **1819**: 97–103

Oh S-J (2005) Arabidopsis CBF3/DREB1A and ABF3 in transgenic rice increased iolerance to abiotic stress without stunting growth. Plant Physiol **138**: 341–351

Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential expression results. Genome Biol **11**: 220

Palaisa KA, Morgante M, Williams M, Rafalski A, Palaisa KA, Morgante M, Williams M, Rafalskiab A (2017) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci published by : American Society of Plant Biologists (ASPB) Linked references are available on JSTOR for this article : Contrasting Effects of. **15**: 1795–1806

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet **40**: 1413–1415

Pedmale U V, Huang SC, Zander M, Cole BJ, Hetzel J, Ljung K, Reis PAB, Sridevi P, Nito K, Nery JR, et al (2017) HHS Public Access. **164**: 233–245

Planes MD, Niñoles R, Rubio L, Bissoli G, Bueso E, García-Sánchez MJ, Alejandro S, Gonzalez-Guzmán M, Hedrich R, Rodriguez PL, et al (2015) A mechanism of growth inhibition by abscisic acid in germinating seeds of Arabidopsis thaliana based on inhibition of plasma membrane H+-ATPase and decreased cytosolic pH, K+, and anions. J Exp Bot **66**: 813–825

Ramanathan V, Feng Y (2009) Air pollution, greenhouse gases and climate change: Global and regional perspectives. Atmos Environ **43**: 37–50

Rawat V, Abdelsamad A, Pietzenuk B, Seymour DK, Koenig D, Weigel D, Pecinka A, Schneeberger K (2015) Improving the annotation of Arabidopsis Lyrata using RNA-Seq data. PLoS One **10**: 1–12

Redekar N, Pilot G, Raboy V, Li S, Saghai Maroof MA (2017) Inference of transcription regulatory network in low phytic acid soybean seeds. Front Plant Sci. doi: 10.3389/fpls.2017.02029

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**: 139–40

Roy S (2016) Function of MYB domain transcription factors in abiotic stress and epigenetic control of stress response in plant genome. Plant Signal Behav **11**: e1117723

Sagor GHM, Zhang S, Kojima S, Simm S, Berberich T, Kusano T (2016) Reducing cytoplasmic polyamine oxidase activity in Arabidopsis increases salt and drought tolerance by reducing reactive oxygen species production and increasing defense gene expression. Front Plant Sci **7**: 1–16

Schlaen RG, Mancini E, Sanchez SE, Perez-Santángelo S, Rugnone ML, Simpson CG,

Brown JWS, Zhang X, Chernomoretz A, Yanovsky MJ (2015) The spliceosome assembly factor GEMIN2 attenuates the effects of temperature on alternative splicing and circadian rhythms. Proc Natl Acad Sci **112**: 9382–9387

Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T, et al (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J **31**: 279–292

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

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

Song L, Huang S -s. C, Wise A, Castanon R, Nery JR, Chen H, Watanabe M, Thomas J, Bar-Joseph Z, Ecker JR (2016) A transcription factor hierarchy defines an environmental stress response network. Science (80-) **354**: aag1550-aag1550

Song Q, Grene R, Heath LS, Li S (2017) Identification of regulatory modules in genome scale transcription regulatory networks. BMC Syst Biol **11**: 140

Stress C, Kreps JA, Wu Y, Chang H, Zhu T, Wang X, Harper JF, Mesa T, Row M, Diego S, et al (2002) Transcriptome changes for Arabidopsis in response to. Society **130**: 2129–2141

Suzuki N, Bassil E, Hamilton JS, Inupakutika MA, Zandalinas SI, Tripathy D, Luo Y, Dion E, Fukui G, Kumazaki A, et al (2016) ABA is required for plant acclimation to a combination of salt and heat stress. PLoS One **11**: 1–21

Suzuki N, Devireddy AR, Inupakutika MA, Baxter A, Miller G, Song L, Shulaev E, Azad RK, Shulaev V, Mittler R (2015) Ultra-fast alterations in mRNA levels uncover multiple players in light stress acclimation in plants. Plant J **84**: 760–772

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

Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T, Blanchette M (2010) Computational analysis of whole-genome differential allelic expression data in human. PLoS Comput Biol **6**: 24

Wang H, Wang H, Shao H, Tang X (2016) Recent advances in utilizing transcription factors to improve plant abiotic stress tolerance by transgenic technology. Front Plant Sci. doi: 10.3389/fpls.2016.00067

Wang L, Si Y, Dedow LK, Shao Y, Liu P, Brutnell TP (2011) A low-cost library construction protocol and data analysis pipeline for illumina-based strand-specific multiplex RNA-seq. PLoS One. doi: 10.1371/journal.pone.0026426

Welti R, Li W, Li M, Sang Y, Biesiada H, Zhou H-E, Rajashekar CB, Williams TD, Wang X (2002) Profiling membrane lipids in plant stress responses. J Biol Chem **277**: 31994–32002

Yadav S, Irfan M, Ahmad A, Hayat S (2011) Causes of salinity and plant manifestations to salt stress: a review. J Environ Biol **32**: 667–85

Yoshida T, Fujita Y, Sayama H, Kidokoro S, Maruyama K, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2010) AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. Plant J **61**: 672–85

Yoshida T, Mogami J, Yamaguchi-Shinozaki K (2014) ABA-dependent and ABA-independent signaling in response to osmotic stress in plants. Curr Opin Plant Biol **21C**: 133–139

You J, Chan Z (2015) ROS regulation during abiotic stress responses in crop plants. Front Plant Sci **6**: 1–15

Zhang W, Ruan J, Ho THD, You Y, Yu T, Quatrano RS (2005) Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana. Bioinformatics **21**: 3074–3081

Zhu JK (2002) Salt and drought stress signal transduction in plants. Annu Rev Plant Biol **53**: 247–273

3.0 APPENDIXES

3.1 APPENDIX 1: SCRIPT FOR MAPPING READS ON GENOME

```
## Mapping reads via STAR

#!/bin/bash

# Building genome index for mapping reads on genome

STAR --runThreadN 16 --runMode genomeGenerate --genomeDir

/groups/songli_lab/RegNetRNAseq/data/Gnmdir --genomeFastaFiles

/groups/songli_lab/RegNetRNAseq/data/TAIR10_Chr.all.fasta --sjdbGTFfile

/groups/songli_lab/RegNetRNAseq/data/Araport11_GFF3_genes_transposons.201606.gtf

# mapping reads via STAR

GNM=/groups/songli_lab/RegNetRNAseq/data/TAIR10_Chr.all.fasta;

GTF=/groups/songli_lab/RegNetRNAseq/data/Araport11_GFF3_genes_transposons.201606.gtf

GnmDir=/groups/songli_lab/RegNetRNAseq/data/Gnmdir;

# setting out the directories;

dir1=/groups/songli_lab/RegNetRNAseq/

echo "$dir1"

cd $dir1/data/SRA;

for f1 in *.sra;

do

    f2=${f1%.sra}

    echo "$f1"

    echo "$f2"

 cd $dir1/data/fastq

 if [ -f $f2"_2.fastq" ]

     then

       f3=$f2"_1.fastq"

       f4=$f2"_2.fastq"

       echo "$f3"
```

```
        echo "$f4"
# cutadapt -q 20 -a AGATCGGAAGAGC -A AGATCGGAAGAGC --minimum-length 30 -o
home/shamima/data/prepro/$f2"_1.p.fastq" -p home/shamima/data/prepro/$f2"_1.fastq.p.fastq" $f3
$f4
cd /$dir1/data/bam
mkdir $f2
echo "-------------------------------------------"
        STAR   --runThreadN 16 \
             --genomeDir $GnmDir \
             --readFilesIn /$dir1/data/fastq/$f3 /$dir1/data/fastq/$f4 \
             --outSAMstrandField intronMotif \
             --outFileNamePrefix /$dir1/data/bam/$f2/$f2 \
             --outSAMtype BAM SortedByCoordinate;
else
cd /$dir1/data/bam
mkdir $f2
echo "#########################################"
         STAR --runThreadN 16 \
           --genomeDir $GnmDir \
           --readFilesIn /$dir1/data/fastq/$f2.fastq \
           --outSAMstrandField intronMotif \
           --outFileNamePrefix /$dir1/data/bam/$f2/$f2 \
           --outSAMtype BAM SortedByCoordinate;
fi
done
```

3.2 APPENDIX 2: SCRIPT FOR COUNTING READS

```
#!/bin/bash

#Readcounting by FeatureCounts

GNM=/groups/songli_lab/RegNetRNAseq/data/TAIR10_Chr.all.fasta;

GTF=/groups/songli_lab/RegNetRNAseq/data/Araport11_GFF3_genes_transposons.201606.gtf

GnmDir=/groups/songli_lab/RegNetRNAseq/data/Gnmdir;

# setting out the directories;

dir1=/groups/songli_lab/RegNetRNAseq/

cd $dir1/data/SRA

  for f1 in *.sra

  do

    f2=${f1%.sra}

    echo "$f1"

    echo "$f2"

    cd $dir1/data/rc1

    mkdir $f2

    cd $dir1/data/fastq

     if [ -d  $f2"_2" ]

    then

      bamsuffix=Aligned.sortedByCoord.out.bam;

      outdir=$dir1/data/rc1

      file1=$f2$bamsuffix
```

```
        aligndir=$dir1/data/bam/$f2
        featureCounts -T 8 \
                -t exon \
                -g gene_id \
                -p \
                -a $GTF \
                -o $outdir/$f2/$f2.readcount.txt \
                 $aligndir/$file1
      else
        bamsuffix=Aligned.sortedByCoord.out.bam;
        outdir=$dir1/data/rc1
        file1=$f2$bamsuffix
        aligndir=$dir1/data/bam/$f2
        featureCounts -T 8 \
                -t exon \
                -g gene_id \
                -a $GTF \
                -o $outdir/$f2/$f2.readcount.txt \
                 $aligndir/$file1
fi
done
```

3.3 APPENDIX 3: SCRIPT FOR MERGING READS AND FPKM

```
#!/Usr/bin/env Rscript
#load library
library(DESeq2)
#Step 0. Set path-----------------------------------------
BasePATH=setwd("~/RNAseqProject")
print(BasePATH)
tmpGSE= list.files(path = BasePATH, pattern = "^GSE")
GseDir= tmpGSE[ !grepl(".R", tmpGSE)]
for (j in GseDir)
{
 print(j)
 SETPATH=paste0(BasePATH,"/",j)
 print(SETPATH)
 setwd(SETPATH)
 getwd()
 #Getting GSEaccession list fronm the directory
 GseNum=gsub("GSE","",j)
 designName=paste0("design_",GseNum,".csv")
 print(designName)
 #Create design matrix-------------------------------------------
 sampleInfo <-read.table(designName, sep=',',as.is=T, header=T)
 print(sampleInfo)
 head(sampleInfo)
 rownames(sampleInfo)<-sampleInfo[,1]

 #Load Data and generate readcountMAtrix--------------------------------
 GSEName=paste0("",j,".csv")
 readcount_table<-(read.table(paste(sampleInfo[1,1]),sep='\t',as.is=T, header=T))
 res <- readcount_table[1];
 colnames(res)[1] <- "GENE ID"
```

```
 Readcountlist<-read.csv(GSEName,header=FALSE)
for (i in 1:nrow(Readcountlist))
{
  print(paste(Readcountlist[i,]))
  readcount_table<-(read.table(paste(Readcountlist[i,]),sep='\t',as.is=T, header=T))
  res<- cbind(res, readcount_table[7])
  colnames(res)[i+1] <- paste(Readcountlist[i,])
}
ReadcountMatrixName<-paste0("",j,"countmatrix.csv")
print(ReadcountMatrixName)
write.csv(as.data.frame(res),file="results/ReadcountMatrixName.csv")
InputDF<-data.frame(res)
print(InputDF)
InputDF<-data.frame(res)
InputDF2<-InputDF[,-1]
rownames(InputDF2)<-InputDF[,1]
print(InputDF2)
#Differential expression analysis
dds = DESeqDataSetFromMatrix(countData = InputDF2, colData=sampleInfo, design =
~Treatment)
dds<-DESeq(dds)
res2<-results(dds)
summary(res2)
#Generating Normalized readcount matrix----------------------
countData  <- data.frame(InputDF2)
dim(countData)
dds.EstSizFac <- estimateSizeFactors(dds)
norm.count=counts(dds.EstSizFac, normalized=TRUE)
head(norm.count)
head(countData)
norm.count.round = round(norm.count, 3)
```

```
 head(norm.count.round)
 #------
NormcountMatrixName<-paste0(j,"Norm.count.csv")
write.csv(as.data.frame(norm.count.round),file="results/NormcountMatrixName.csv", row.names=T)
#Getting FPKM and load library edgeR--------------------
library(edgeR)
head(norm.count)
GeneidLengthFile="SRR.ID.Length"
AnnoData  <- read.table(GeneidLengthFile, sep='\t',as.is=T, header=T)
head(AnnoData)
head(AnnoData)
norm.count.DGEList <- DGEList(counts=norm.count, genes=AnnoData[,c("Geneid","Length")])
print(norm.count.DGEList)
norm.rpkm <- rpkm(norm.count.DGEList, norm.count.DGEList$genes$Length)
head(norm.rpkm)
norm.rpkm.round = round(norm.rpkm, 3)
head(norm.rpkm.round)
NormFPKMName<-paste0("",j,"Norm.FPKM.csv")
write.csv(as.data.frame(norm.rpkm.round),file="results/NormFPKMName.csv", row.names=T)
# Getting average FPKM value for each GSE------------------------
FPKM.j<-read.csv("results/NormFPKMName.csv",header=T)
WholeAveFPKM=data.frame(GeneID=res[,1])
head(WholeAveFPKM)
for (i in unique(sampleInfo[,2]))
{
   print(i)
   SRR<-sampleInfo[(which(sampleInfo[,2]==i)),]
   print(SRR)
   FPKM.condition<-FPKM.j[,colnames(FPKM.j) %in% SRR[,1]]
   head(FPKM.condition)
   MeanFPKMCondition<-as.matrix(rowMeans(FPKM.condition))
```

```
    head(MeanFPKMCondition)
    colnames(MeanFPKMCondition)<-i
    head(MeanFPKMCondition)
    WholeAveFPKM=cbind(WholeAveFPKM,MeanFPKMCondition)
}
head(WholeAveFPKM)
MeanFPKMName<-paste0("",j,"Mean.FPKM.csv")
write.csv(as.data.frame(WholeAveFPKM),file="results/MeanFPKMName.csv", row.names=T)
```

3.4 APPENDIX 4: RSCRIPT FOR DIFFERENTIAL EXPRESSION ANALYSIS

```
#Loading data
#!/Usr/bin/env Rscript
#Step 1. Load Library
# load library
library(DESeq2)
#!/Usr/bin/env Rscript
#Setup working directory
#setwd("/groups/songli_lab/RegNetRNAseq/data/GSE_List/GSE_RC_text/GSE59699")
#Step 2. Setup your experimental design.
#set up row names
#create "Design Matrix"
# Step 3. Load data
Readcountlist<-read.csv("GSE64870.csv", header=FALSE)
readcount_table<-(read.table(paste(Readcountlist[1,]),sep='\t',as.is=T, header=T))
res <- readcount_table[1];
colnames(res)[1] <- "GENE ID"
for (i in 1:nrow(Readcountlist))
{
  print(paste(Readcountlist[i,]))
  readcount_table<-(read.table(paste(Readcountlist[i,]),sep='\t',as.is=T, header=T))
  res<- cbind(res, readcount_table[7])
  colnames(res)[i+1] <- paste(Readcountlist[i,])
}
write.csv(as.data.frame(res),file="results/GSE64870_countMatrix.csv")
sampleInfo <-read.table("design_64870.csv", sep=',',as.is=T, header=T)
print(sampleInfo)
```

```
head(sampleInfo)
rownames(sampleInfo)<-sampleInfo[,1]
InputDF<-data.frame(res)
print(InputDF)
InputDF<-data.frame(res)
InputDF2<-InputDF[,-1]
rownames(InputDF2)<-InputDF[,1]
print(InputDF2)
# Step 4. Creat DESeq2 object
dds = DESeqDataSetFromMatrix(countData = InputDF2, colData=sampleInfo, design = ~Treatment)
# Step 5. perform differential expression analysis
dds<-DESeq(dds)
res2<-results(dds)
# Step 6. check out the results
summary(res2)
plotMA(res2, main="DESeq2", ylim=c(-2,2))
plotCounts(dds, gene=which.min(res$padj), intgroup="condition")
# Step 7. write output
resSig <- subset(res2, res2$padj < 0.1)
resSigOrdered <- resSig[order(resSig$padj),]
write.csv(as.data.frame(resSigOrdered),file="results/GSE_64870_DESeq2.csv")
```