UNDERSTANDING PLANT STRESS RESPONSES: USING SYSTEMS BIOLOGY APPROACH AND TEXT MINING METHODS

by

Rita Kusi-Appiah Hayford

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Bioinformatics Data Science

Summer 2022

UNDERSTANDING PLANT STRESS RESPONSES: USING SYSTEMS BIOLOGY APPROACH AND TEXT MINING METHODS

by

Rita Kusi-Appiah Hayford

Approved: _____
Cathy H. Wu, Ph.D.
Chair of Center for Bioinformatics & Computational Biology

Approved: _____
Levi T. Thompson, Ph.D.
Dean of the College of Engineering

Approved: _____
Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Cathy H. Wu, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Venu (Kal) Kalavacharla, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Cecilia Arighi, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Jyothi Thimmapuram, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

Thanks to the Almighty God for giving me the strength and ability to understand, learn and complete this journey. This dissertation was jointly supervised by Dr. Cathy Wu of the University of Delaware and Dr. Venu (Kal) Kalavacharla of Delaware State University (DSU)/USDA-NIFA. Foremost, I would like to express my sincere gratitude to my advisor, Dr. Cathy Wu, for accepting me to her lab. I am truly thankful to Dr. Wu for her time, effort, enthusiasm, motivation, and expert guidance to aid in the completion of the dissertation. My most profound appreciation to Dr. Kal, this chapter of my life would not have been possible without his guidance. Dr. Kal welcomed me to the Molecular Genetics and Epigenomics (MGE) Lab in 2012 as a master's student and has supported me ever since. Thank you, Dr. Kal, for believing in me! A debt of gratitude is also owed to Dr. Cecilia Arighi, who I have worked closely with since joining Dr. Wu's lab. Dr. Arighi has always been patient in explaining the scientific concepts. She provides valuable feedback, and her critical views have helped shape this research. I would like to express my heartfelt gratitude to Dr. Jyothi Thimmapuram of the Bioinformatics Core at Purdue University. I worked with Dr. Thimmapuram before my Ph.D. Studies. As a committee member, I am incredibly thankful for your collaborating support, time, and thoughtful insights into my research. I thank all the lovely members of the MGE lab and Dr. Wu's lab for the friendly working environment. I take the opportunity to thank all my co-authors in the publication presented for the excellent team spirit. My special thanks to the administrative team, Dr. Karen Hoober and, Ms. Andrea Trungold, for their support in

making my studies at UD comfortable. The administrative assistance of Ms. Laurieann Phalen of DSU cannot go unrecognized.

My biggest thanks to my family and friends far and near for their support during my Ph.D. Studies. Many thanks go to my friends Sachin Gavali, Pelisa Antoh, Prince Boakye, and Harold Abaidoo-Ayin for always being there.

To my husband Samuel Hayford, I would not have completed this work without your unfailing love and support. You have been amazing and always cheered me on, and for this, I say thank you. To my children, Josie, Ivan, Nell, and Kelsie, you are my inspiration for achieving greatness; your understanding has gotten me this far.

Last but not least, I would like to thank my mom, Ms. Cecilia Frimpong, and my siblings for their continuous encouragement, prayers, and love. To my father, Mr. Kingsley Kusi-Appiah, of blessed memory, I thank you for instilling the spirit of endurance and perseverance in me. It was your greatest dream for me to become a Ph.D., but unfortunately, you are not around to witness this moment. I did it for you, daddy, and I am dedicating this dissertation to you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The world's population is growing exponentially, with a current growth rate of approximately 1.1% per year. As of 2017, the number of undernourished people in the world was estimated as 821 million (FAO). Climate variability is increasingly viewed as a significant cause of hunger. Due to climate change and global warming, different biotic and abiotic stresses pose a severe threat to the agricultural sector limiting crop productivity worldwide. In the natural environment, plants face multiple biotic and abiotic stresses and the combined effect of these stresses has a tremendous impact on crop yield. In this regard, it is important to take steps for a genome-scale molecular understanding of stress response mechanisms in plants to help develop stress-tolerant cultivars. The amount of scientific literature on plant stress responses keeps increasing and this could pose a challenge to researchers as important information could be buried in the text. Biologists need to obtain a comprehensive knowledge of biological systems. For this reason, an approach to combine our knowledge in 'omics' studies and text mining to link genes to their function in plants when imposed with environmental stress has been implemented. The overarching objective of this dissertation is to improve our understanding of stress response in plants using 'omics' technologies and to complement standard enrichment analysis with text mining methods.

- First, RNA-Seq approach was used to understand the molecular mechanisms underlying stress response in an important bioenergy crop switchgrass (*Panicum virgatum* L.). Switchgrass was exposed to a single drought (DT) treatment and combinations of DT and heat (HT) (DTHT) stress treatment at different times points. Unique and overlapping genes and pathways were identified in response to DT and combined DTHT stress.
- Secondly, we established a pipeline to automatically retrieve information on plant stress from the scientific literature to support the annotation of switchgrass. This pipeline integrates data from relevant resources to efficiently retrieve publications to study stress response in plants. The data collected is stored in MongoDB and used to predict additional role of the stress-responsive genes in switchgrass from the first study. We validated a candidate gene, Phenylalanine ammonia-lyase 1, involved in stress response in switchgrass. A preliminary work was conducted by evaluating in-house and publicly available tools to build a pipeline to retrieve literature to study stress response in the model plant Arabidopsis.
- Lastly, to support the enrichment analysis performed in the first study, we created and visualized a functionally organized group of terms and pathways using ClueGO. The differentially expressed genes (DEGs) of the switchgrass transcriptome data was uploaded into ClueGO, a plug-in of Cytoscape software. ClueGO integrates files from Gene Ontology,

KEGG and Reactome, they were used to perform a ClueGO network of terms and pathways.

The approach of combining systems biology and text mining methods to study stress response has generated valuable data to complement existing knowledge on plant stress. Such knowledge will eventually be useful to create a resource for the plant biology community and help with crop improvement in the long term.

Chapter 1

INTRODUCTION AND BACKGROUND

Plants (*Viridiplantae* in Latin) are living organisms of the kingdom Plantae. Plantae includes all lands plants: mosses, ferns, vines, herbs, bushes, trees, conifers, flowering plants, and green algae. Plants are one of the significant groups of living organisms required for the function of the biosphere. There are over 300,000 plant species identified on earth. As autotrophs, plants can make their food through photosynthesis. The oxygen released by plants in the same process promotes aerobic life. Essential foods produced by plants include carbohydrates, fats, and proteins, and it will be impossible to have most life on earth without these food sources. Plants uptake carbon dioxide during photosynthesis which helps to reduce the greenhouse effect and climate change [1]. Besides food, humans depend on plants for their basic needs, such as clothing, shelter, and medicine. Plants are essential to the ecosystem they occupy and contribute to improving the habitat by filtering the air, water and soil. Due to the growing world population, the basic needs of humans are also increasing. Bennett (2010) asserted that humans obtain 85% of their calories from 20 different plant species while 60% of their calories are obtained from three plant species *Oryza sativa* (rice), *Triticum aestivum* (wheat), and *Zea mays* (maize) [2,3].

The world's population is predicted to exceed nine billion by 2050, with a 95% certainty that by 2050 the expected population growth will be between 9.4 and 10.1 billion [4,5]. This increase in population is of significant interest to various disciplines in Agriculture and the food production industry. This means the amount of food

produced would need to be doubled to meet the nutritional demand of the growing population. One of the significant challenges in the 21st century is the issue of the changing climate and extreme weather conditions making it one of the utmost importance in research communities and among interested stakeholders . The risks and challenges of climate change in addition to the use of land for biofuel production, also reduce the production of crops and supply of food. Abiotic and biotic stresses severely affect crops production globally with average yield loss of more than 70% [6]. Abiotic stresses greatly influence plant growth and yield; they include both physical and chemical factors such as DT, HT, cold, salinity, UV-B light intensities, flooding, nutrient deficiencies, and gas emissions. These abiotic stresses have been extensively studied, and it is estimated that abiotic stresses decrease the yield of major crops ranging from 50 to 70% [7]. Similarly, a recent report revealed the loss of crop yields worldwide up to 51-82% from the effect of DT, extreme temperatures, deficiency, and toxicity of nutrients [8]. Biotic stresses, including pathogens, bacteria, fungi, viruses, nematodes, and pests, invade plants causing vast economic losses [9,10]. A previous survey on major crops showed that pathogens, insects, pests, and weeds caused an average yield loss of potatoes ranging from 17.2% to 30% in rice, 21.4% in soybean, and 22.5% in maize [11].

Under natural conditions, plants are exposed to combinations of two or more stresses such as DT and salinity, salinity and HT, and combinations of DT and HT. The damage becomes even more deleterious due to the co-occurrence of multiple abiotic stresses or the interaction of multiple abiotic and biotics stresses. Previous reports indicate the influence of high temperatures on the spread of pests and pathogens. Furthermore, many abiotic stresses have been shown to reduce the defense

mechanisms of plants rendering them susceptible to pathogen infection [12]. A study which compared all the major US weather disasters between the period of 1980 and 2012 identified that a combination of DT and HT stress caused severe agricultural losses of about $200 billion. However, over the same period of the study, the effect of only DT on agricultural production was $50 billion, suggesting that the presence of the second stress can increase the impact of the first [10].

Being sessile organisms plants are unable to escape biotic and abiotic stresses. As a result, they have evolved to live in environments where they are usually exposed to these stress factors. Plants sense these stresses and produce enormous molecular, biochemical, physiological, and morphological responses. Additionally, the regulatory or transcriptional machinery of plants become activated during stress and eventually generate an appropriate response. Whereas some of the responses produced by plants to different stress conditions can be general; being commonly manifested irrespective of the type of stress applied , others can be specific to a particular stress. Many studies have reported the transcriptome changes of plants to single stresses[13,14]. Despite the effort of studying the effect of individual stress, the complex interactions between multiple biotic and abiotic stresses have been under explored. Therefore understanding the mechanisms of how plants respond to single and combinations of stresses is therefore crucial in developing a broad-spectrum stress-tolerant crops [8,15]. Recent studies have shown that the molecular response of plants to multiple stresses is unique and cannot be directly inferred from the response of plants to to the individual stresses applied separately. The simultaneous occurrence of biotic and abiotic stresses adds a degree of complexity since the responses are usually regulated by various hormone signalling pathways that may inhibit or interact one another [16].

As such, plant scientist uses systems biology approach to understand the complex biological system of plants by focusing on their highly interconnected components. As a biology-based interdisciplinary field, system biology focuses on complex interactions among different components in the biological system. The interconnected components of a self-sustaining unit functioning together is referred as a system [17]. A system in the biological world could refer to a biological ecosystem, organisms, organ system, tissue system, cellular system, genes, proteins and metabolites. For the functional sustainability of the system, the member components need to function together. System biology offers a comprehensive view of plant systems, utilizing a holistic approach by integrating the molecular data at various hierarchical levels. It is an approach whereby a system of interacting units is analyzed as a whole rather than analyzing its individual members separately. The knowledge and research in system biology has increased over the last decade. Systems biology has been evolving as a promising tool to study stress responses and adaptation. The rapid progress in high-throughput data generation has provided the platform for multi-omics systems biology research, offering answers to complex issues by enabling virtual test and analysis and hypothesis testing. The core datasets of systems biology are; transcriptomics, proteomics, and metabolomics, providing the expression levels of transcripts, proteins, and metabolites. Omics technology and bioinformatics are important to understanding the molecular systems underlying plants' function. A significant aspect of systems biology is network analysis which provides a platform for omics data visualization. Data visualization helps to reduce the intrinsic complexity of the data [18–21].

Several microarray and transcriptome experiments have been used to study the transcriptional response pattern(s) to different biotic and abiotic stresses. These experiments have mostly been conducted in the model plant *Arabidopsis thaliana*. Models in biological science are organisms with huge amount of existing biological information that makes them conducive as examples for other species. Models are usually less complex and easy to use them for experiments. Among the other plant models available, Arabidopsis is the most widely studied 'reference system' for all biological processes by the plant science community. Over the last decade Arabidopsis has emerged as the primary experimental system for essentially all aspects of plant biology. In addition, because of the close evolutionary relationships between all flowering plants, discoveries in Arabidopsis have been readily translated to other plant species such as economically important crops. Interestingly, discoveries made in Arabidopsis have impacted research in human biology. Almost all major research breakthroughs in plant science over the last 20 years have relied on development of Arabidopsis as a reference system. Arabidopsis has a small genome size (~132 Mbp) with available complete and annotated sequenced genome. Moreover, Arabidopsis has a rapid life cycle and has efficient transformation methods utilizing Agrobacterium tumefaciens [22]. As of September 2012, 23913 people and 9968 institutes/groups registered as Arabidopsis researchers in the Arabidopsis Information Resource (TAIR) [23].

Current omics technology has provided a good platform to conduct various studies to understand the molecular mechanisms underlying stress response in plants. This has produced an enormous amount of data such as transcriptome data in plants during stress conditions. There are a number of resources developed with data on

stress response genes in plants. This information keeps increasing making it a challenge for researchers and curators to keep up to date. The relevant information hidden in the text needs to be extracted automatically without being labor intensive in less time. Text mining helps to answer specific research questions, it filters a large amount of research and extracts the relevant information. It can identify and match patterns and trends across millions of articles which is helpful in determining additional research that is needed to answer research questions. Additionally, text mining helps to draw inferences by combining information from multiple sources [24]. Text mining applications which include information extraction system aim to extract facts reported in the biology literature [25,26]. Significant efforts have been made to develop methods to extract relation between biological entities (e.g. genes-disease or drug-disease relation) [27]. The three main methods discussed for relation extraction are; co-occurrence-based, rule-based and machine learning. In the co-occurrence-based, the co-reference of the bioentities in the same sentence or paragraph indicates a relation [28,29]. In the rule-based method of relation extraction, automatic programs encoded with linguistics and or biological knowledge (such as plain-text, syntactic pattern, specific words linked to the biological relation) are implemented to extract relations from text [25,30]. Classification of entities have been used in the machine learning method of relation extraction. Naïve Bayes [31] and Support Vector Machine [32] are examples of machine learning methods that have been used in relation extraction. Co-occurrence-based method have been widely used for example in extracting mutation-disease relation [33]. We established a pipeline that integrates relevant resources to find signicant data to study plant stress. We use the co-occurrence method of extraction to identify stress genes and important keywords and

sentences describing the role of the genes from the literature. Since a n number of plant resources is targeted on the reference plant *Arabidopsis thaliana* compared to non-model crops such as switchgrass our goal is to use the data collected with the pipeline to study stress-responsive genes in switchgrass.

Switchgrass is of interest due to its high biomass and its ability to maintain growth and development with small amounts of water. It grows in marginal areas and has a potential to be used as a biofuel crop. Switchgrass is classified as a weedy or invasive warm season plant that grows throughout North America [34]. Drought is a serious abiotic threat to the sustainability of the switchgrass, especially under the current paradigms of changing climate across the globe. Similarly, various studies have reported the impact of high temperatures on switchgrass, emphasizing physiology, cell wall composition, biomass, and yield.

Like many plants, switchgrass is impacted by multiple environmental stresses. However, there have been no high-throughput studies (transcriptome analysis) of switchgrass during single DT and DT and HT stress combinations. Thus, the unique molecular response by switchgrass to combination of abiotic stresses has been underexplored. Exploring these using a system biology approach can elucidate new knowledge about stress-responsive genes in not just switchgrass but plants in general. We generated gene expression data on switchgrass when exposed to single DT and combination of DTHT stress. Bioinformatics tools and network analysis resources were utilized for GO enrichment, co-expression, pathway and gene network analysis. Since reports on adequate knowledge of genes to support crop production and adaptation have primarily focused on a small number of well-studied model plants , one of the objectives was to predict the role of the switchgrass stress-genes by

homology. To support the annotation of the stress-responsive genes, a pipeline integrating text mining methods was established to automatically retrieve stress genes from literature and link them to their function or biological process. The role of a candidate stress gene in the phenylpropanoid pathway of switchgrass was experimentally validated using proteomics and bioinformatics methods. The overall goal of this study is to improve our understanding on plant stress response using systems biology and text mining methods. This study combines high throughput technologies, bioinformatics tools, and text mining to extract stress-responsive genes and their relationship with biological processes. This approach is useful to develop resources to help with the sustainability of improving switchgrass ecotypes to tolerate multiple stresses. Understanding plants response to adverse environmental conditions and importantly crops that can withstand multiple stresses is crucial for sustainable food security.

REFERENCES

[1] Fernando 2012,I,1-5;doi:10.3390/plants1010001 n.d.

[2] Bennett BC. Plants as Food, Economic Botany n.d.

[3] Maroyi A. Ethnobotanical study of wild and cultivated vegetables in the eastern cape province, south africa. Biodiversitas 2020;21:3982–8. https://doi.org/10.13057/biodiv/d210908.

[4] Molotoks A, Smith P, Dawson TP. Impacts of land use, population, and climate change on global food security. Food Energy Secur 2021;10:1–20. https://doi.org/10.1002/fes3.261.

[5] YORK UNN (UNDESA). World population prospects 2019. 2019.

[6] Raza A, Razzaq A, Mehmood SS, Zou X, Zhang X, Lv Y, et al. Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. Plants 2019. https://doi.org/10.3390/plants8020034.

[7] Francini A, Sebastiani L. Abiotic stress effects on performance of horticultural crops. Horticulturae 2019;5. https://doi.org/10.3390/horticulturae5040067.

[8] Teshome DT, Zharare GE, Naidoo S. The Threat of the Combined Effect of Biotic and Abiotic Stress Factors in Forestry Under a Changing Climate. Front Plant Sci 2020;11. https://doi.org/10.3389/fpls.2020.601009.

[9] Peterson RKD, Higley LG. Biotic stress and yield loss. 2000. https://doi.org/10.1201/9781420040753.

[10] Suzuki N, Rivero RM, Shulaev V, Blumwald E, Mittler R. Abiotic and biotic stress combinations. New Phytol 2014;203. https://doi.org/10.1111/nph.12797.

[11] Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. The global burden of pathogens and pests on major food crops. Nat Ecol Evol 2019;3:430–9. https://doi.org/10.1038/s41559-018-0793-y.

[12] Atkinson NJ, Lilley CJ, Urwin PE. Identification of genes involved in the response of arabidopsis to simultaneous biotic and abiotic stresses. Plant Physiol 2013;162:2028–41. https://doi.org/10.1104/pp.113.222372.

[13] Suzuki N, Rivero RM, Shulaev V, Blumwald E, Mittler R. Abiotic and biotic stress combinations. New Phytol 2014;203:32–43. https://doi.org/10.1111/nph.12797.

[14] Rasmussen S, Barah P, Suarez-Rodriguez MC, Bressendorff S, Friis P, Costantino P, et al. Transcriptome Responses to Combinations of Stresses in Arabidopsis. Plant Physiol 2013;161:1783–94. https://doi.org/10.1104/pp.112.210773.

[15] Zandalinas SI, Mittler R, Balfagón D, Arbona V, Gómez-Cadenas A. Plant adaptations to the combination of drought and high temperatures. Physiol Plant 2018;162. https://doi.org/10.1111/ppl.12540.

[16] Asselbergh B, De Vleesschauwer D, Höfte M. Global switches and fine-tuning-ABA modulates plant pathogen defense. Mol Plant-Microbe Interact 2008;21:709–19. https://doi.org/10.1094/MPMI-21-6-0709.

[17] Trewavas A. A brief history of systems biology. Plant Cell 2006;18:2420–30. https://doi.org/10.1105/tpc.106.042267.

[18] Sahoo JP, Behera L, Sharma SS, Praveena J, Nayak SK, Samal KC. Omics Studies and Systems Biology Perspective towards Abiotic Stress Response in Plants. Am J Plant Sci 2020;11:2172–94. https://doi.org/10.4236/ajps.2020.1112152.

[19] Mochida K, Shinozaki K. Advances in omics and bioinformatics tools for systems analyses of plant functions. Plant Cell Physiol 2011;52:2017–38. https://doi.org/10.1093/pcp/pcr153.

[20] Jamil IN, Remali J, Azizan KA, Nor Muhammad NA, Arita M, Goh HH, et al. Systematic Multi-Omics Integration (MOI) Approach in Plant Systems Biology. Front Plant Sci 2020;11. https://doi.org/10.3389/fpls.2020.00944.

[21] Chawla K, Barah P, Kuiper M, Bones AM. Systems Biology: A Promising Tool to Study Abiotic Stress Responses. Omi Plant Abiotic Stress Toler 2011:163–72. https://doi.org/10.2174/978160805092511101010163.

[22] Benfeyp. Microsoft Word - Report of the Workshop 2_10 Final.doc 2008:1–12.

[23] Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Res 2012. https://doi.org/10.1093/nar/gkr1090.

[24] Krallinger M, Valencia A, Hirschman L. Linking genes to literature: Text mining, information extraction, and retrieval applications for biology. Genome Biol 2008. https://doi.org/10.1186/gb-2008-9-s2-s8.

[25] Torii M, Arighi CN, Li G, Wang Q, Wu CH, Vijay-Shanker K. RLIMS-P 2.0: A generalizable rule-based information extraction system for literature mining of protein phosphorylation information. IEEE/ACM Trans Comput Biol Bioinforma 2015;12:17–29. https://doi.org/10.1109/TCBB.2014.2372765.

[26] Van Landeghem S, De Bodt S, Drebert ZJ, Inzé D, Van De Peer Y. The potential of text mining in data integration and network biology for plant research: A case study on Arabidopsis. Plant Cell 2013. https://doi.org/10.1105/tpc.112.108753.

[27] Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research. BMC Bioinformatics 2015;16:1–17. https://doi.org/10.1186/s12859-015-0472-9.

[28] Alako BTF, Veldhoven A, van Baal S, Jelier R, Verhoeven S, Rullmann T, et al. CoPub Mapper: Mining MEDLINE based on search term co-publication. BMC Bioinformatics 2005;6:1–15. https://doi.org/10.1186/1471-2105-6-51.

[29] Junge A, Jensen LJ. CoCoScore: Context-aware co-occurrence scoring for text mining applications using distant supervision. Bioinformatics 2020;36:264–71. https://doi.org/10.1093/bioinformatics/btz490.

[30] Li G, Ross KE, Arighi CN, Peng Y, Wu CH, Vijay-Shanker K. miRTex: A Text Mining System for miRNA-Gene Relation Extraction. PLoS Comput Biol 2015;11:1–24. https://doi.org/10.1371/journal.pcbi.1004391.

[31] Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. A primer on learning in Bayesian networks for computational biology. PLoS Comput Biol 2007. https://doi.org/10.1371/journal.pcbi.0030129.

[32] Hong G. LNAI 3651 - Relation Extraction Using Support Vector Machine 2014. https://doi.org/10.1007/11562214.

[33] Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, Peterson T, et al. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. Bioinformatics 2011;27:408–15. https://doi.org/10.1093/bioinformatics/btq667.

[34] Ayyappan V, Saha MC, Thimmapuram J, Sripathi VR, Bhide KP, Fiedler E, et al. Comparative transcriptome profiling of upland (VS16) and lowland (AP13) ecotypes of switchgrass. Plant Cell Rep 2017. https://doi.org/10.1007/s00299-016-2065-0.

Chapter 2

GLOBAL ANALYSIS OF SWITCHGRASS (*PANICUM VIRGATUM* L.) TRANSCRIPTOMES IN RESPONSE TO INTERACTIVE EFFECTS OF DROUGHT AND HEAT STRESSES

Hayford, R.K., Serba, D.D., Xie, S. *et al.* Global analysis of switchgrass (*Panicum virgatum* L.) transcriptomes in response to interactive effects of drought and heat stresses. *BMC Plant Biol* **22,** 107 (2022). (https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-022-03477-0)

2.1 Abstract

Background: Sustainable production of high-quality feedstock has been of great interest in bioenergy research. Despite the economic importance, high temperatures and water deficit are limiting factors for the successful cultivation of switchgrass in semi-arid areas. There are limited reports on the molecular basis of combined abiotic stress tolerance in switchgrass, particularly the combination of drought and heat stress. We used transcriptomic approaches to elucidate the changes in the response of switchgrass to drought and high temperature simultaneously. **Results:** We conducted solely drought treatment in switchgrass plant Alamo AP13 by withholding water after 45 days of growing. For the combination of drought and heat effect, heat treatment (35 °C/25 °C day/night) was imposed after 72 h of the initiation of drought. Samples were collected at 0 h, 72 h, 96 h, 120 h, 144 h, and 168 h after treatment imposition, total RNA was extracted, and RNA-Seq conducted. Out of a total of 32,190 genes, we identified 3,912, as drought (DT) responsive genes, 2,339 and 4,635 as , heat (HT) and drought and heat (DTHT) responsive genes, respectively.

There were 209, 106, and 220 transcription factors (TFs) differentially expressed under DT, HT and DTHT respectively. Gene ontology annotation identified the metabolic process as the significant term enriched in DTHT genes. Other biological processes identified in DTHT responsive genes included: response to water, photosynthesis, oxidation-reduction processes, and response to stress. KEGG pathway enrichment analysis on DT and DTHT responsive genes revealed that TFs and genes controlling phenylpropanoid pathways were important for individual as well as combined stress response. For example, hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (HCT) from the phenylpropanoid pathway was induced by single DT and combinations of DTHT stress. **Conclusion:** Through RNA-Seq analysis, we have identified unique and overlapping genes in response to DT and combined DTHT stress in switchgrass. The combination of DT and HT stress may affect the photosynthetic machinery and phenylpropanoid pathway of switchgrass which negatively impacts lignin synthesis and biomass production of switchgrass. The biological function of genes identified particularly in response to DTHT stress could further be confirmed by techniques such as single point mutation or RNAi.

2.2 Background

Plants in the field are exposed to various environmental stresses which affect production and yield. These environmental stresses include abiotic factors such as DT, HT, and salinity and biotic stresses like pathogens, and insect pests, [1]. Abiotic stresses are reported to reduce about 50% of crop production [2]. Stress tolerance research has primarily focused on the response of plants to individual stress with limited information on plants' adaptability to combined stresses such as HT and DT

and salinity and DT [3–5]. Moreover, plants exhibit a unique expression pattern when exposed to multiple stresses [6]. Hence to bridge the knowledge gap, we have compared the transcriptional response of switchgrass when exposed to individual DT stress or a combination of DT and HT stresses. The combined effect of DT and HT stresses has been shown to cause more damage to plants than when these stresses occur at separate times [7,8]. The mechanisms used by plants to adapt to multiple stresses can be complex. It has been shown that the effect of one stress could have a synergistic or antagonizing effect on other stress. DT, salinity, high and low temperature have been shown to promote the occurrence of pathogens and pests [5]. In addition, the antagonizing effect of cold stress on osmotic stress during the induction of dehydration-responsive gene *RD29A* has been reported [9]. Abscisic acid (ABA) was found to antagonize jasmonate-ethylene signaling pathways and mediates defense gene expression and disease resistance in Arabidopsis [10]. Multiple stress in plants led to the expression of common overlapping genes due to a cross-talk of a signaling pathway. A previous study identified 22 genes that were induced commonly during DT, cold, and NaCl treatment [11]. Some of the molecular mechanisms adopted by plants to combat stress include the release of HT shock proteins or chaperons that are expressed during exposure to environmental cues [12].

Transcriptome analysis of Arabidopsis showed that HT resistance is conferred by HT stress-responsive genes, plant hormones, and antioxidant enzymes [13]. The importance of transcriptional gene regulation in plants under DT and HT stresses has been previously reported [13]. RNA sequencing (RNA-Seq) has been commonly used to identify genome-wide transcript profiles in plants. Stress-responsive genes have been identified in tobacco and Arabidopsis when exposed to combined DT and HT

stress by RNA-Seq technology [14,15]. Plant responses to single stress treatment of cold, high light, salt, HT, and flagellin have been compared to various combinations of these six pair of stresses (cold and high light, salt and HT, salt and high light, HT and high light, HT and flagellin respectively). The outcome of this study revealed how plants have evolved to withstand combination of these stresses [4]. The combined effect of DT and HT stress has been studied in wheat [16]. The effect of combined abiotic stress signaling such as DT, salinity, and metal in rice was found to be complex with the involvement of multiple genes, differential expression patterns in different developmental tissues, and protein-protein interaction [17]. Furthermore, the separate impact of DT and HT and their combined effect on grain filling, physiological, vegetative, and yield traits were investigated in wheat [8].

Switchgrass (*Panicum virgatum* L.) is a C4 warm-season perennial grass identified as a potential bioenergy crop [18,19]. It has been investigated for lignocellulosic ethanol production in the US, Canada, and Europe [20] due to its high biomass yield. It serves as a potential alternative to nonrenewable fossil fuels, thereby providing energy security sources [21]. Switchgrass requires a minimal amount of water and nutrients and can grow on marginal croplands [22]. Its rapid growth rate and broad adaptability contribute to a stable and high biomass supply. Switchgrass positively impacts the soil by improving soil quality, preventing erosion, and reducing soil nutrients [23].

Switchgrass, like many other plants, is generally faced with extreme biotic and abiotic stresses. These stresses can be detrimental by causing retardation in plant growth, development, and even death [24]. DT is a significant abiotic stress that limits switchgrass use as biofuel production. There is evidence of DT as an essential

economic risk factor affecting biofuel production [25]. Molecular mechanisms underlying DT responses in plants have been addressed in various articles [26]. A previous report suggests DT could considerably reduce the yield and quality of biomass for biofuel production [27]. The effect and response of switchgrass germplasms to DT stress have been evaluated in previous studies [28–30]. High temperatures in the Southern United States are projected to reduce switchgrass biomass in 2080-2090 [22]. Similarly, various studies have reported the impact of high temperatures on switchgrass, emphasizing physiology, cell wall composition, biomass, and yield. A significant decrease in biomass yield was observed across various switchgrass genotypes due to the impact of high temperatures [22,31]. There is increasing research in switchgrass, and among the area of research is gene regulation. Transcriptome analysis has been used to determine genes associated with biomass production in switchgrass [29]. The characterization of DT and HT responsive microRNAs has been recently reported [18]. Besides, the role of microRNAs during DT and salt stress in switchgrass has been reported [32].

Although switchgrass is an essential bioenergy crop, less information on the biology of switchgrass is available when imposed with abiotic stresses [23]. The molecular mechanisms of the tolerance of switchgrass to hot and dry climates is not well studied [18]. Therefore, understanding the effect of stress combinations in switchgrass will be important to reveal genes associated with important traits such as biomass and biofuel production in response to multiple environmental stresses. Additionally, breeding DT and HT resistant switchgrass cultivars will be an important milestone. Although several studies have reported switchgrass response to a single DT or HT stress, there are no reports as far as we know on the combination of DT and HT

abiotic stresses in switchgrass, especially with prolonged exposure to DT and HT stresses.

To better understand plant responses to the full complement of environmental stresses, it is important to compare data on single stresses with data on multiple stresses. It is also important to identify the early transcriptional response to DT and HT stress versus the prolonged exposure of switchgrass to these stresses. This will provide an idea of signaling cross-talk in systems biology [33]. In this study, we used RNA-Seq approach to characterize and quantify gene expressions in response to DT and combined effects of DT and HT stresses in switchgrass.

2.3 Results

2.3.1 RNA-Seq data quality and summary

A total of 6,965 million paired-end reads were obtained from RNA-Seq samples. The number of reads in each sample was 129 million on average. Around 85% of the reads can be aligned to the reference genome. About 63% of reads were aligned to genic regions. To assess the similarities and differences among these samples, we performed a hierarchical cluster analysis of the RNA-Seq data (**Figure 2.1**). We found that non-treated samples were grouped together except the 72 h DT treated samples. In the group of stress treated samples, DTHT samples were grouped together except 144 h DTHT sample, which clustered with the group of DT samples.

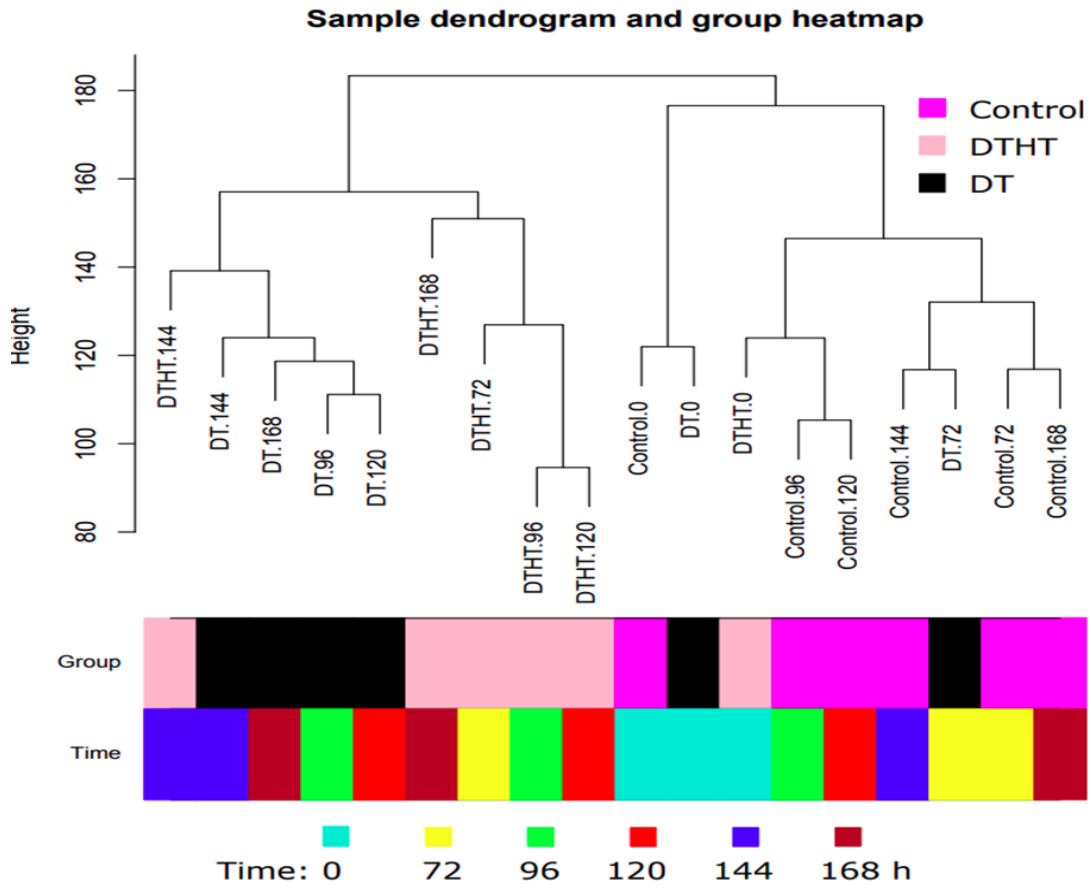


Figure 2.1**:** Hierarchical clustering analysis of Control, DT, and DTHT treated samples.

2.3.2 Analysis of DT and DTHT responsive genes in switchgrass

From the analysis, many genes were identified in response to the DTHT compared to only DT stress. In total, 3,912 out of 32,190 genes were identified as DT and 4,635 as DTHT responsive genes. Among those, 1,615 genes were shared between the DT and DTHT data sets, when DT samples were compared to plants exposed to combined DTHT stress. These commonly expressed genes likely play critical roles in DT and HT tolerance in switchgrass. Further analysis showed that 1,432 out of 2,282 of the up-regulated responsive genes were unique(Figure 2.2A) and 1,604 out of

2,345 down-regulated genes were unique to DTHT (Figure 2.2B). Similarly, for DT samples, 1,307 out of 2,157 up-regulated responsive genes were unique, while 1,013 out of 1,754 down-regulated genes were unique(Figure 2.2A and 2.2B).

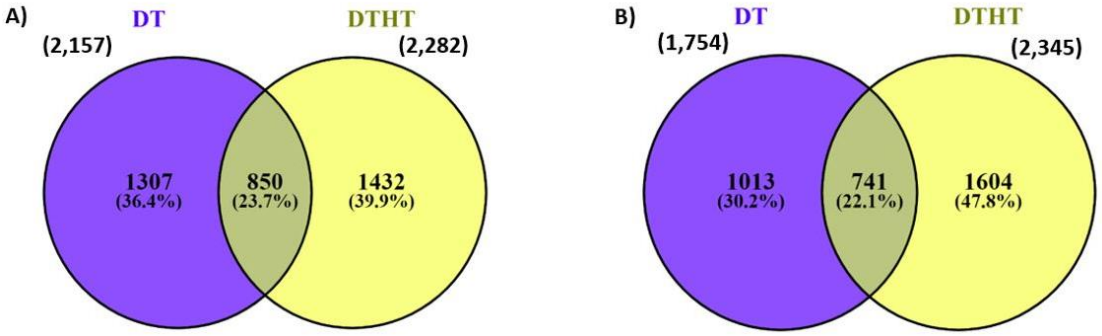


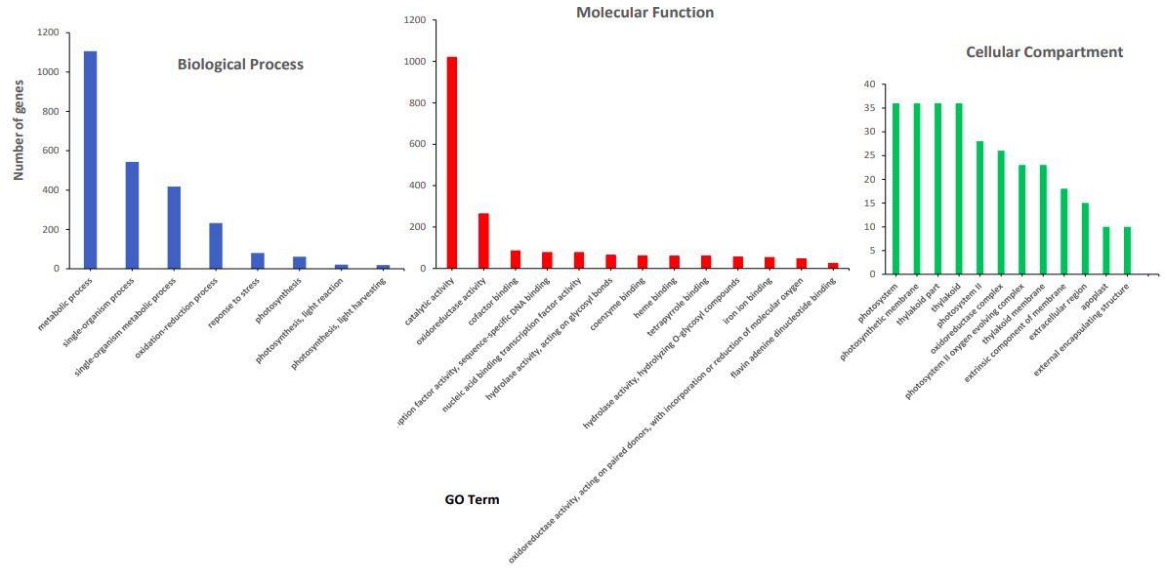
Figure 2.2: The number of common and specific up-regulated (A), and down-regulated (B) genes among switchgrass during DT and DTHT stress in the Venn diagram. The genes were significantly differentially expressed (DE) in more than one comparison of the time point, 0 h, 72 h, 96 h, 120 h, 144 h, and 168 h. DE genes for each comparison were quantified at log2 fold changes and P-value <0.05.

In our data, Pavir.6 KG130600.v4.1 provided the best hit to Arabidopsis AT1G22360.1 (UDP-glucosyl transferase 85A2 (UGT85A2) and it is the only DT-responsive gene that showed both up and down-regulation between the time points after imposing DT treatment (**Additional file 5, DT**). This gene was significantly down-regulated at time points DT 96 h and DT 120 h after which its expression markedly up-regulated at 168 h.

Through GO enrichment analysis (**Figure 2.3a, 2.3b, Additional file 6**), we found that there were significantly enriched terms in all biological process, molecular

function, and cellular component functional categories. In the biological process category, the enriched GO terms included photosynthesis, single-organism metabolic process, and metabolic process. GO enrichment analysis show that the GO terms; "response to stress" and "response to water", with p-values (0.00042 and 0.00054, respectively) were smaller than 0.05 although the FDR values were above 0.05 (0.083 and 0.093, respectively). Eight out of 15,902 genes belonged
to the GO term of response to water in the switchgrass genome whereas seven out of 3,912 DT responsive genes also belonged to the GO term of response to water. In molecular function, some of the enriched terms were oxidoreductase activity, catalytic activity, and cofactor binding. In the cellular component category, the enriched terms were photosystem, photosynthetic membrane, and thylakoid part. We further performed KEGG enrichment analysis on the DT responsive genes. We found that these DEGs were enriched in the following KEGG pathways (**Additional file 7**): protein phosphatase 2C, glutaredoxin 3, homeobox−leucine zipper protein, jasmonate ZIM domain−containing protein, and solute carrier family, xyloglucan: xyloglucosyl transferase, HSP20 family protein, adenylate kinase, and UDP-glucuronate decarboxylase.

a)


Biological Process
Molecular Function
Cellular Compartment
Number of genes
GO Term

b)

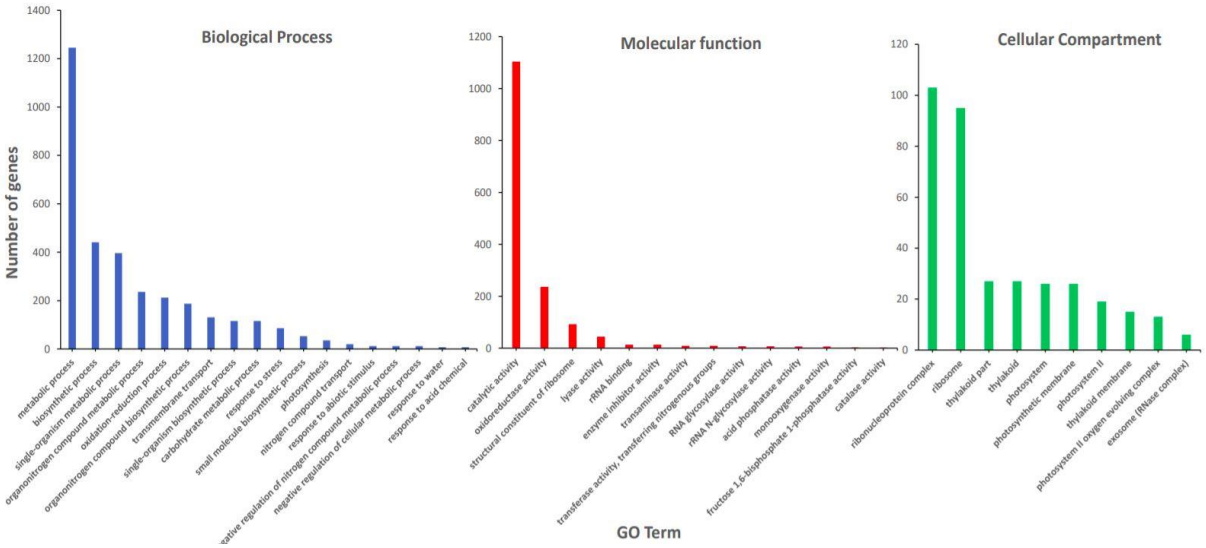

Biological Process
Molecular function
Cellular Compartment
Number of genes
GO Term

Figure 2.3: **a**. The Gene Ontology (GO) terms enriched by responsive genes to DT stress. The DEGs were annotated against the GO database. The GO terms are in the three GO domains (biological process, molecular function and cellular compartment). These terms were significantly enriched ($p < 0.05$) in DT treated samples compared to control plants. The number of genes enriched in each term were plotted against the GO term. **b**. The Gene Ontology (GO) terms enriched by responsive genes to DTHT stress. The DEGs were annotated against the GO database. The GO terms are in the three GO domains (biological process, molecular function, and cellular compartment). These terms were significantly enriched ($p < 0.05$) in combined DT and HT treated samples compared to control plants. The number of genes enriched in each term were plotted against the GO term.

Pavir.9NG755000.v4.1 which provided the best hit to (ATHCHIB, B-CHI, CHI-B, HCHIB, PR-3, PR3) is a basic chitinase gene was significantly down-regulated at 144/72 h and subsequently up-regulated after prolonged DT and HT stress at 168/96 h. Similarly, genes such as Pavir.5KG627200.v4.1, Pavir.2NG348700.v4.1 and Pavir.2NG348700.v4.1 with best hit to Arabidopsis genes encoding delta 1-pyrroline-5-carboxylate synthase 2 (AT3G55610.1), cytochrome P450, family 76, subfamily C (AT2G45550.1), polypeptide 4, and DUF1012 (AT5G43745.1) respectively were significantly down-regulated at 144/72 h (**Additional file 5, DTHT**). These genes at 168/96 h were significantly up-regulated after prolonged DT and HT stress, suggesting the possible role of these genes in protecting the plant during extreme environmental conditions

To study the functions of these responsive genes, GO enrichment analysis was performed. The main GO term from the enrichment analysis was the GO term (GO:0008152; metabolic process) which showed significant enrichment (FDR; 0.0014) (**Figure 2.3).** None of the GO terms shows significant enrichment in combined DT and HT stress responsive genes, indicating that DTHT transcriptomic

changes were not predictable from single stress treatments. In the category of biological process, there were 10 most enriched GO terms with P-value <= 0.05. These 10 GO terms were response to water, single-organism metabolic process, single-organism biosynthetic process, response to abiotic stimulus, organonitrogen compound metabolic process, photosynthesis, oxidation-reduction process, response to stress, nitrogen compound transport, and transmembrane transport respectively. We further performed KEGG enrichment analysis on the DTHT responsive genes (4,635 genes). We found that these responsive genes were enriched in the following KEGG pathways (**Additional File 7**): adenylate kinase and protein phosphatase 2C.

2.3.3 HT responsive genes in switchgrass

The HT stress genes were deduced from the DEGs of DT and DTHT. In total, 2,338 out of 32,190 genes were identified as HT responsive genes (**Additional file 5**). There were 1,064 up-regulated genes and 1,274 down-regulated genes. The functions of these responsive genes and GO annotation were presented (**Additional file 6**). In the category of biological process, these genes showed enrichment in the GO terms of organic cyclic compound catabolic process, organonitrogen compound catabolic process and heterocycle catabolic process, etc. In the category of molecular function, these genes showed enrichment in the GO terms of organic cyclic compound catabolic process, organonitrogen compound catabolic process and heterocycle catabolic process, etc. In the categories of cellular components, these genes showed enrichment in the GO terms of photosystem II oxygen-evolving complex, photosystem II, and thylakoid membrane. We also performed KEGG enrichment analysis on the HT

specific responsive genes. We found that these responsive genes were enriched in the jasmonate ZIM domain-containing protein pathway (**Additional File 7**).

2.3.4 Transcription factors (TF) for DT, DTHT and HT responses

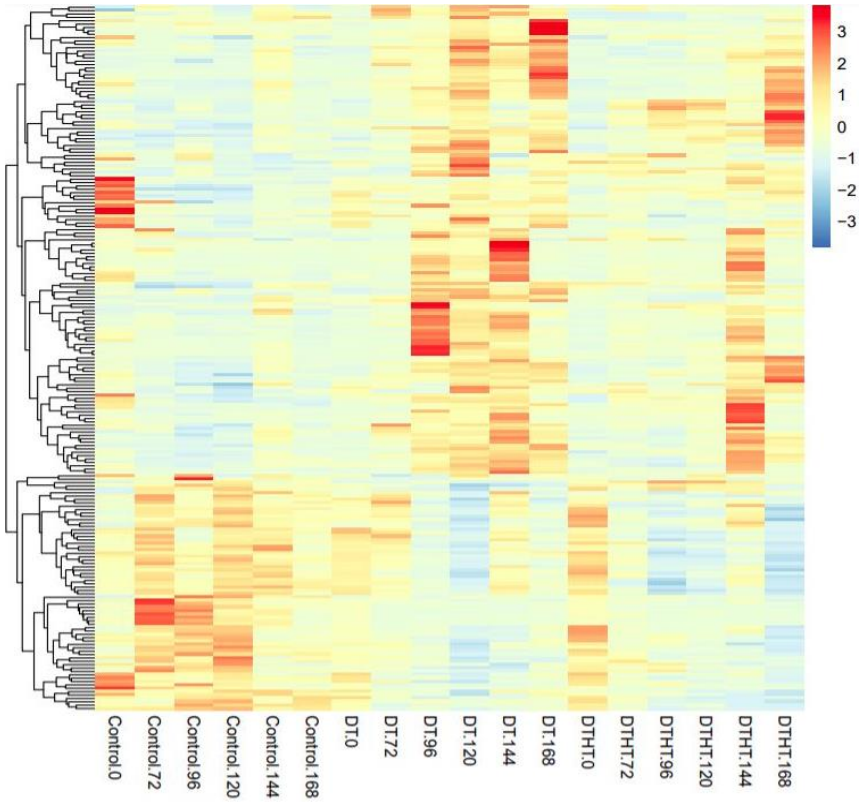
The TFs identified from the analysis are shown in **Table 1,** and **Additional file 8**. These DT and DTHT responsive TFs belong to 45 different TF families. Out of 91,838 proteins on the switchgrass genome, 3,948 were identified as transcription factors (TFs). A total of 1,383 TFs were identified out of 32,190 genes that were used for identifying stress responsive genes. There were 209 genes identified as TFs out of 3,912 DT responsive genes. Heat maps were generated to show expression patterns of these 209 genes in all the samples (**Figure 2.4A**). Similarly, there were 220 genes identified as TFs out of 4,635 DTHT responsive genes. A heat map was generated to show expression patterns of these 220 genes in all the samples (**Figure 2.4).** A total of 106 genes out of the 2,339 predicted HT responsive genes, were identified as TFs. Heat map was generated to show expression patterns of these 106 genes in all the samples (**Figure 2.4C**).

Table 2.1**.** Different families of TFs responsive to solely DT and combined DTHT stresses.

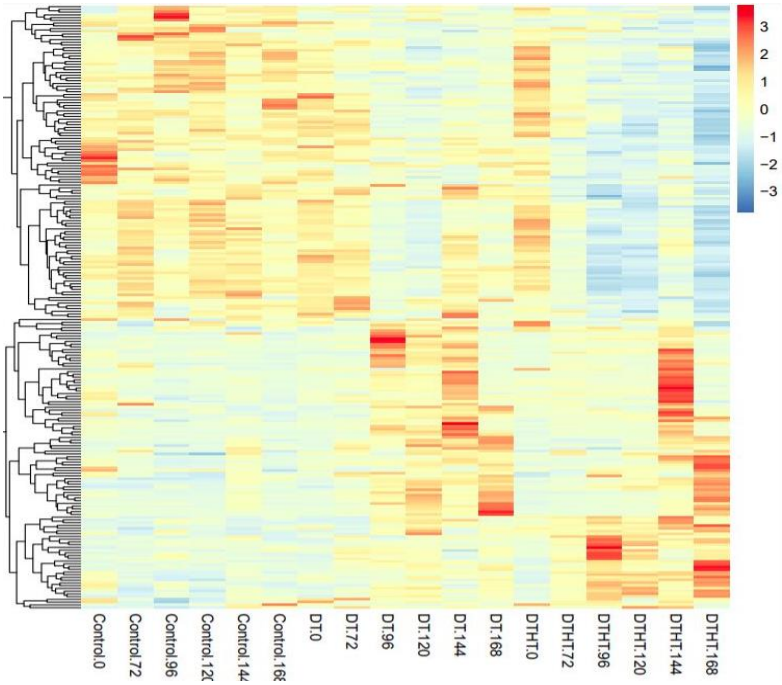
Transcription factor type	DTvsCtrl	DTHTvsCtrl	DTHTvsDT
bHLH	22	20	10
NAC	16	15	13
ERF	19	19	6
bZIP	17	17	5
MYB_related	10	17	10
MYB	12	15	7
WRKY	14	11	6
HD-ZIP	15	7	3

C3H	7	13	1

a)


3
2
1
0
−1
−2
−3
Control.0
Control.72
Control.96
Control.120
Control.144
Control.168
DT.0
DT.72
DT.96
DT.120
DT.144
DT.168
DTHT.0
DTHT.72
DTHT.96
DTHT.120
DTHT.144
DTHT.168

b)


3
2
1
0
−1
−2
−3
Control.0
Control.72
Control.96
Control.12
Control.14
Control.16
DT.0
DT.72
DT.96
DT.120
DT.144
DT.168
DTHT.0
DTHT.72
DTHT.96
DTHT.120
DTHT.144
DTHT.168

c)

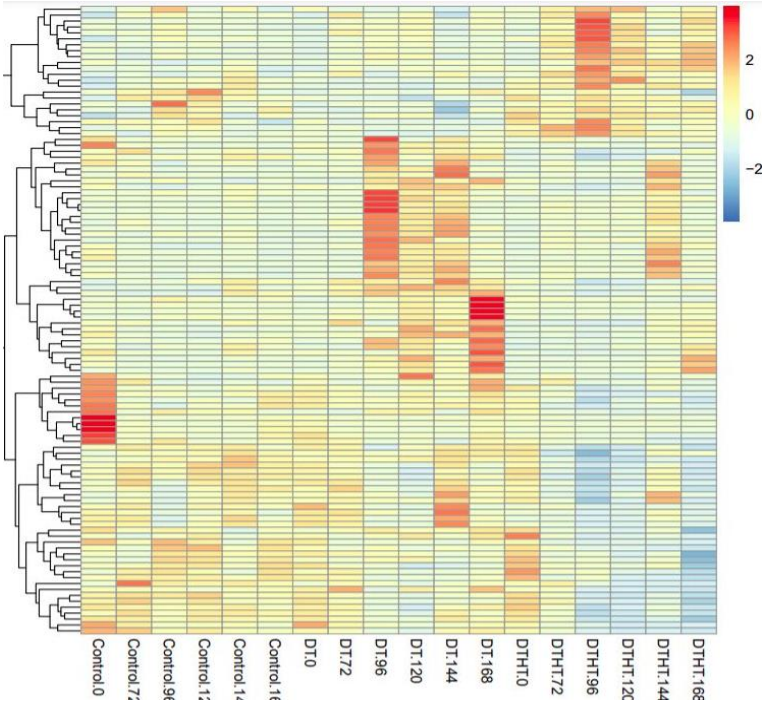
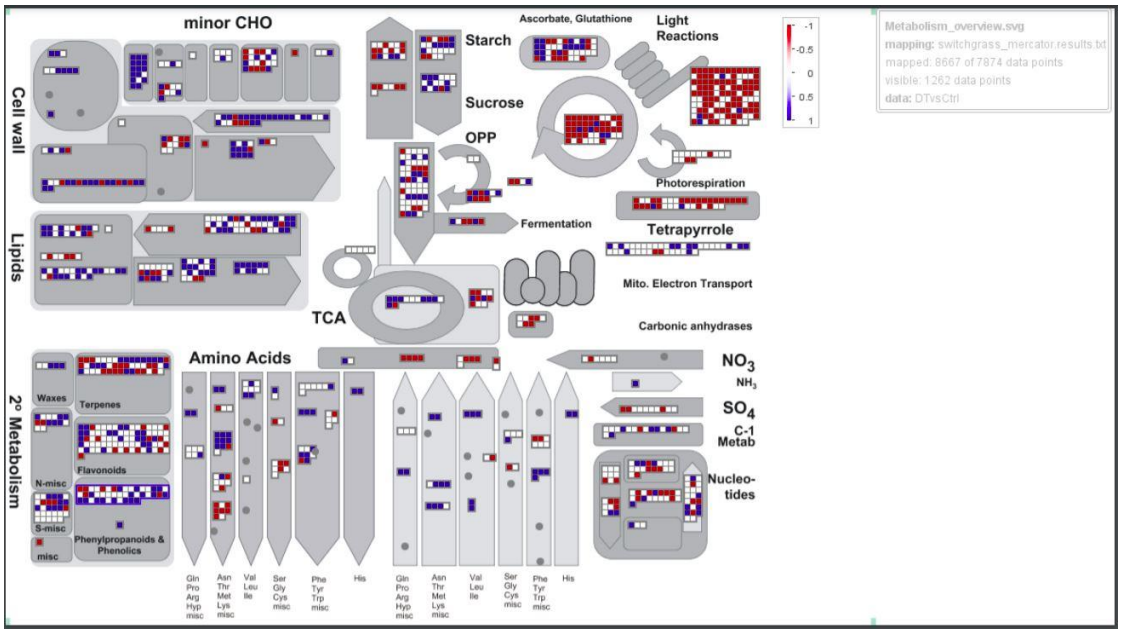

2
0
−2
DT.0
DT.72
DT.96
DT.120
DT.144
DT.168
DTHT.0
DTHT.72
DTHT.96

Figure 2.4: Heat map with clusters based on FPKM values for **A**) DT vs Control, **B**) DTHT vs control and **C**) DTHT vs DT TFs. The Heat map shows a grouping of control samples and stress samples. Extended periods of DTHT to stress samples showed abundant up-regulated TFs (**A** and **B**) and down-regulated TFs (**C**) compared to their control samples. For example, there were more responsive TFs which were up-regulated at time 144/72 h compared to its control sample at Control 144/72 h (**A**)

2.3.5 Pathway analysis of DT and HT responsive genes

An overview of the secondary metabolism pathway is displayed in Figure 2.5 (A and B). We found a large number of plant secondary metabolites such as flavonoids, terpenes, and phenylpropanoids were down-regulated in DTHT vs control samples compared to DT vs control samples.

a)



b)

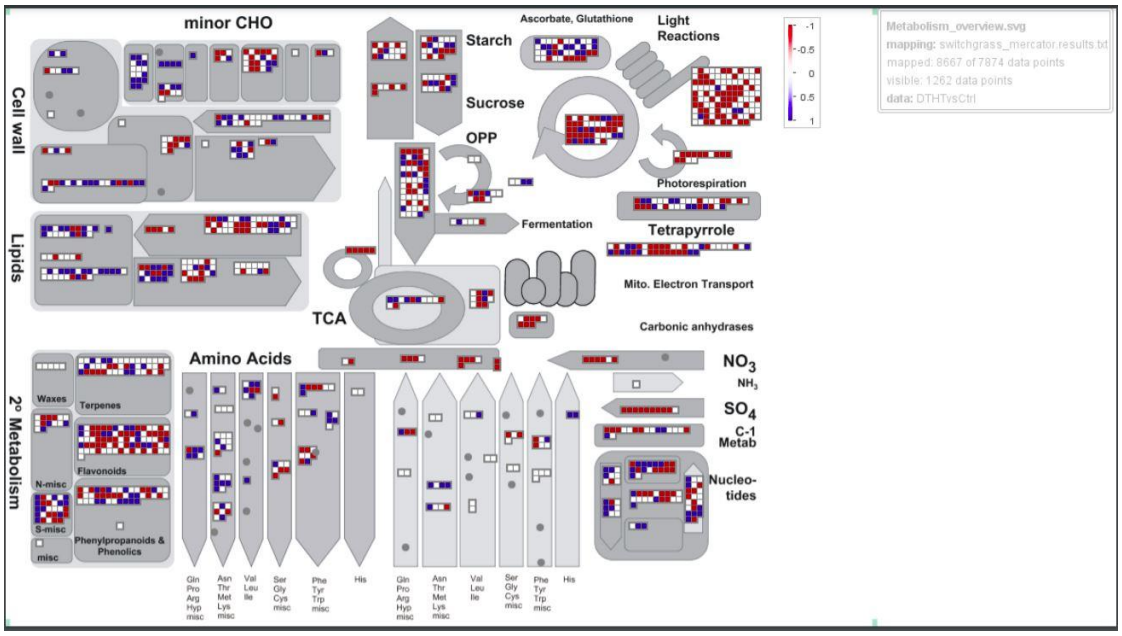


Figure 2.5: Metabolism overview in MapMan showing the DEGs between DT vs Control (A) and DTHT vs control (B) switchgrass samples. The log-fold ratio is indicated as a gradient with red color (down-regulated) and blue color (up-regulation).

2.3.6 Co-expression network

We performed weighted gene co-expression network analysis to identify genes involved in response to the DT and DTHT stresses. Most of co-expressed genes usually participate in the same biological processes [34–36]. In our co-expression analysis, we identified 68 modules with distinct expression patterns (**Additional file 11**). To study whether the DEGs were enriched in some of the modules, Fisher's exact test and multiple test correction (Benjamini-Hochberg method) were performed [4] . Of the modules that have more than 100 genes, DT responsive genes were enriched in module 5, 7, 14, 17 and 25. DTHT responsive genes were enriched in module 1, 2, 3, 7 and 17. HT responsive genes were enriched in module 1, 2, 8, 9, 15, 16, 17 and 25. GO enrichment analysis of the genes of these modules were performed using agriGO.

Results for GO enrichment are provided in (**Additional file 12**). Heat maps were generated for these 12 unique modules (**Additional file 2**). In module 7 and module 17, both DT responsive genes and DTHT responsive genes were enriched. In module 7 and module 17, genes were up-regulated after stress treatment. In module 7, the genes were enriched in GO terms of response to water, response to acid chemical, lipid biosynthetic process, and response to the oxygen-containing compound, or biological process. In module 17, the genes were enriched in GO terms of regulation of nucleic acid-templated transcription, regulation of RNA biosynthetic process, regulation of RNA metabolic process and regulation of transcription, DNA-templated, etc. for biological process. In module 1 (**Figure 2.6**), most genes were up-regulated during the initial HT treatment at DTHT 96/24h. Downregulation of most of the genes in the same module occur and then up-regulated again at an extensive HT at DTHT168/96h. Similarly, in module 8 which is enriched with HT responsive genes, showed upregulation of genes at the initial stage of imposing HT at DTHT96/24h. In module 1, the genes were enriched in GO term biological processes such as translation, peptide biosynthetic process, amide biosynthetic process and peptide metabolic process. In module 8, the genes are enriched in GO terms including; multi-organism reproductive process, multi-multicellular organism process, cell recognition, and pollination for biological processes. Also, DTHT responsive genes were enriched in module 9 with most of the responsive genes recorded at time point DTHT96/24h and DTHT120/48h. A number of the genes recorded at DTHT96/24h and DTHT120/48h were enriched in different class of metabolic processes.

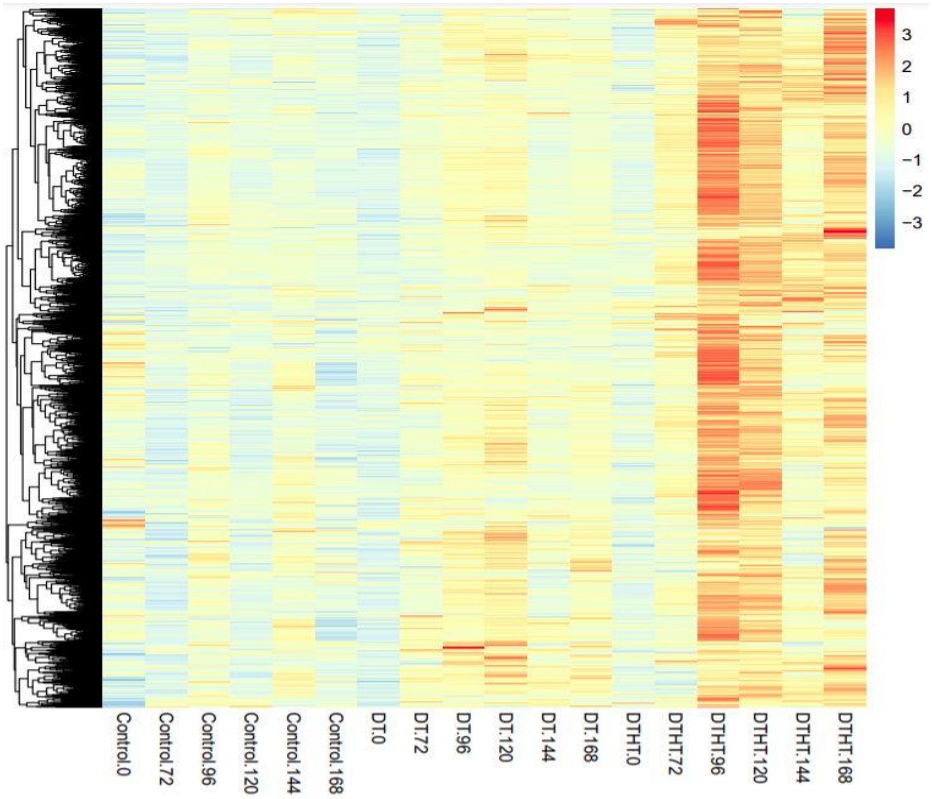


Figure 2.6: Heat map indicating genes enriched in module 1 from the WGCNA analysis. DTHT and HT responsive genes were enriched in module 1

2.3.7 DT and DTHT responsive genes in DroughtDB

There were 386 genes from the switchgrass genome that have the best hits to Arabidopsis genes in the droughtDB [37] Of these 386 genes, 172 were found in the 32,190 genes in this study. Detailed gene expression patterns of these 172 genes were shown in the heat map (**Additional file 3**). Out of these 172 genes, there were 35 DT responsive genes and 27 DTHT responsive genes in which 12 were common (**Additional file 13**). A list of the DT and DTHT genes have been indicated in Table 2.2 and 2.3, respectively. The gene IDs, biological functions, the phenotype of mutants, references, tags of the genes from Arabidopsis can be obtained. For example,

three genes are described in detail which play an important role in DT response: Pavir.1KG544600.v4.1 is homologous to KAT2 in Arabidopsis. In Arabidopsis, the *kat2-3* mutant shows ABA-insensitive phenotypes and KAT2-overexpressing transgenic lines show strong ABA-hypersensitive phenotypes (ABA-induced stomatal closure and inhibition of stomatal opening) [26]. In our data, Pavir.1KG544600.v4.1 showed increased gene expression levels under both DT and DTHT treatments. In Arabidopsis, HAB1/PP2C is known as a major negative regulator of ABA signaling and its mutant shows hypersensitive to ABA [38]. In our data, Pavir.8NG117400.v4.1, homologous to HAB1/PP2C, showed increased gene expression level under both DT and DTHT treatments. Additionally, the ABCG22 (Pavir.9NG742000.v4.1) from Arabidopsis is an ABC-transporter and a knockout of ABCG22 caused Arabidopsis to be more susceptible to DT stress [39]. From our data Pavir.9NG742000.v4.1 showed increased gene expression level under both DT and DTHT treatment. The 386 switchgrass genes with best hits to Arabidopsis genes in droughtDB were used to generate the heat map (**Additional file 3**).

Table 2.2**:** List of DT-responsive genes identified in switchgrass in the droughtDB**.**

Gene id	Gene	Biological Function
Pavir.9NG610900.v4.1	GolS1	Galactinol Synthase, catalyzes the first step in the biosynthesis of Raffinose Family Oligosaccharides (RFOs) from UDP-galactose
Pavir.6NG274900.v4.1	AREB1	bZIP TF, ABRE binding
Pavir.6KG307800.v4.1	ABF4	ABA responsive element (ABRE) binding bZIP factor
Pavir.5KG406700.v4.1	ABCG40	ABC-transporter, ABA import
Pavir.2KG548500.v4.1	OST1/SRK2E	Kinase-like (open stomata 1), activated by ABA, activates SLAC1
Pavir.2NG401700.v4.1	ATHB6	homeodomain protein, target of ABI1
Pavir.2NG618000.v4.1	GolS2	Galactinol Synthase, catalyzes the first step in the biosynthesis of Raffinose Family Oligosaccharides (RFOs) from UDP-galactose
Pavir.9KG306600.v4.1	GSTU17	glutathion s-transferase U17
Pavir.2NG248100.v4.1	MYB44	MYB type TF
Pavir.7KG296100.v4.1	AGO1	Argonaute1
Pavir.9KG354500.v4.1	MYC2	transcriptional activator of ABA signaling
Pavir.4KG090000.v4.1	AVP1	vacuolar membrane H+-Pyrophosphatase
Pavir.9KG421700.v4.1	GolS1	Galactinol Synthase, catalyzes the first step in the biosynthesis of Raffinose Family Oligosaccharides (RFOs) from UDP-galactose
Pavir.6KG279400.v4.1	FAD8	fatty acid desaturase
Pavir.1KG544600.v4.1	KAT2	3-ketoacyl-CoA thiloase-2
Pavir.1KG312700.v4.1	ERD1	chloroplast-targeted Clp protease reg SU
Pavir.3KG112200.v4.1	DHAR2	dehydroascorbate reductase
Pavir.6NG207900.v4.1	XERICO	small protein, N-term- TM domain and RING-H2 zinc-finger motif
Pavir.1NG392600.v4.1	PIP1;4	PIP
Pavir.1NG081300.v4.1	AVP1	vacuolar membrane H+-Pyrophosphatase
Pavir.8NG117400.v4.1	HAB1	PP2C
Pavir.6KG334900.v4.1	XERICO	small protein, N-term- TM domain and RING-H2 zinc-finger motif
Pavir.2KG570400.v4.1	GolS2	Galactinol Synthase, catalyzes the first step in the biosynthesis of Raffinose Family Oligosaccharides (RFOs) from UDP-galactose
Pavir.5KG405500.v4.1	HAB1	PP2C
Pavir.9KG536300.v4.1	SQE1	squalene epoxidase1
Pavir.J678200.v4.1	AAO3	Arabidopsis aldehyde oxidase, catalyzes final step in ABA biosynthesis
Pavir.9KG308600.v4.1	GSTU17	glutathion s-transferase U17

Table 2.3**:** List of genes responsive to combined DT and HT stress in switchgrass from the droughtDB

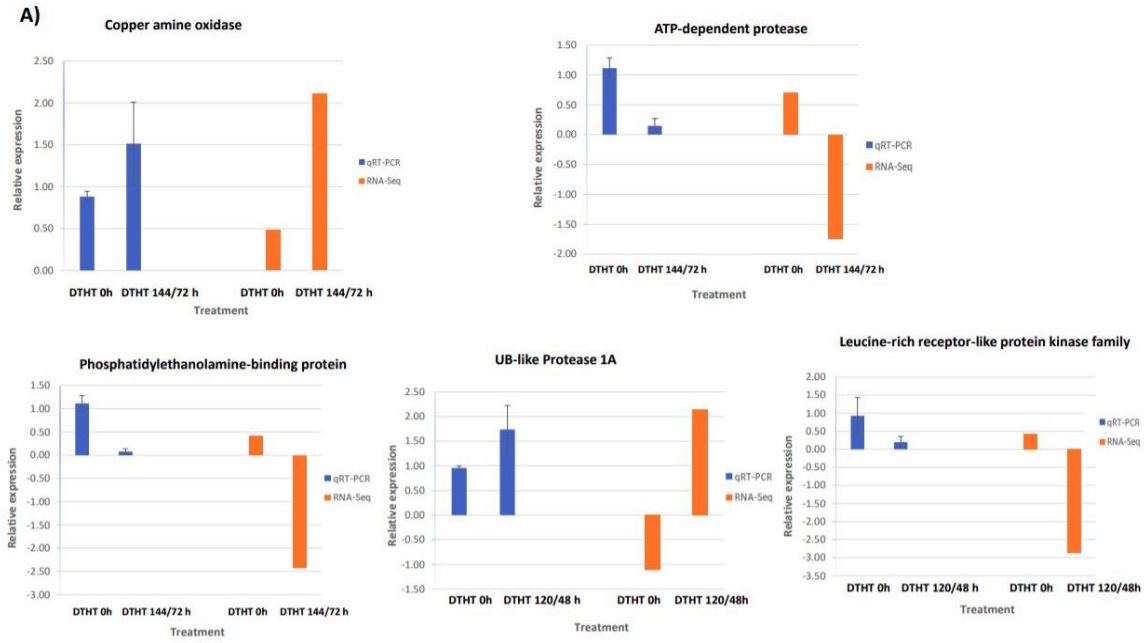
Gene_id	Gene	Biological Function
Pavir.9NG211300.v4.1	ABO1/ELO1	subunit of Elongator, a multifunctional complex with roles in transcription elongation, secretion and tRNA modification
Pavir.9NG493600.v4.1	GSTU17	glutathion s-transferase U17
Pavir.2NG618000.v4.1	GolS2	Galactinol Synthase, catalyzes the first step in the biosynthesis of Raffinose Family Oligosaccharides (RFOs) from UDP-galactose
Pavir.5NG017000.v4.1	SLAH3	guard cell S-type anion channel (SLAC1 homolog)
Pavir.7KG296100.v4.1	AGO1	Argonaute1
Pavir.1NG551600.v4.1	PIP1;4	PIP
Pavir.9NG671400.v4.1	PEPCK	PEP carboxykinase
Pavir.6KG279400.v4.1	FAD8	fatty acid desaturase
Pavir.9KG480900.v4.1	APX2	Ascorbate peroxidase 2, H2O2 scavenger
Pavir.1KG544600.v4.1	KAT2	3-ketoacyl-CoA thiloase-2
Pavir.1KG312700.v4.1	ERD1	chloroplast-targeted Clp protease reg SU
Pavir.3KG112200.v4.1	DHAR2	dehydroascorbate reductase
Pavir.1NG545200.v4.1	AGO1	Argonaute1
Pavir.9NG719800.v4.1	GPA1	alpha subunit of heterotrimeric GTP-binding protein
Pavir.J075500.v4.1	AAO3	Arabidopsis aldehyde oxidase, catalyzes final step in ABA biosynthesis
Pavir.9KG118700.v4.1	GSTU17	glutathion s-transferase U17
Pavir.1NG081300.v4.1	AVP1	vacuolar membrane H+-Pyrophosphatase
Pavir.8NG117400.v4.1	HAB1	PP2C
Pavir.9NG671500.v4.1	PEPCK	PEP carboxykinase
Pavir.6KG334900.v4.1	XERICO	small protein, N-term- TM domain and RING-H2 zinc-finger motif
Pavir.2KG570400.v4.1	GolS2	Galactinol Synthase, catalyzes the first step in the biosynthesis of Raffinose Family Oligosaccharides (RFOs) from UDP-galactose
Pavir.7NG063700.v4.1	MRP4	multidrug resistance-associated protein, ABC transporter
Pavir.9KG517100.v4.1	PEPCK	PEP carboxykinase

Pavir.3KG456000.v4.1	ERD1	chloroplast-targeted Clp protease reg SU
Pavir.6NG268500.v4.1	XERICO	small protein, N-term- TM domain and RING-H2 zinc-finger motif
Pavir.7KG292400.v4.1	CBF4	DREB family TF
Pavir.2KG247300.v4.1	PARP1	poly(ADP-ribose) polymerase

2.3.8 Validation of RNA-Seq results using qRT-PCR

Seven candidate genes responsive to DTHT stress were selected from the RNA-Seq data for validation by performing qRT-PCR (**Figure 2.7A and 2.7B**). The expression pattern of the selected genes was consistent with the RNA-Seq results.

A)


Copper amine oxidase
Relative expression
qRT-PCR
RNA-Seq
DTHT 0h
DTHT 144/72 h
Treatment
ATP-dependent protease
Phosphatidylethanolamine-binding protein
UB-like Protease 1A
DTHT 120/48 h
DTHT 120/48h
Leucine-rich receptor-like protein kinase family

B)

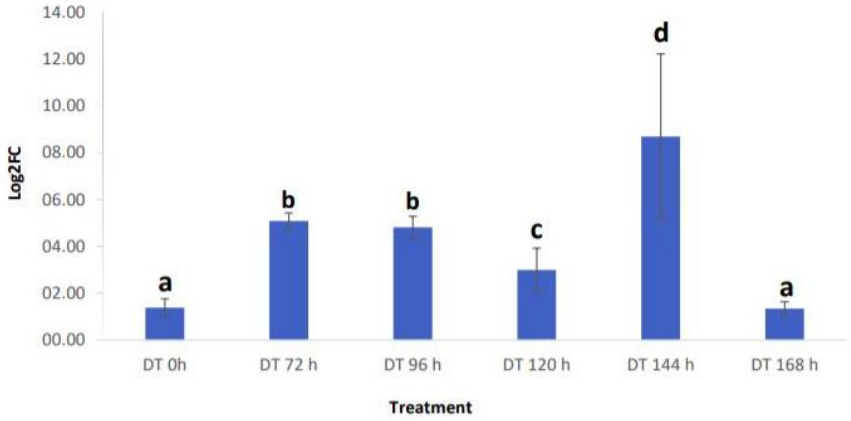

Log2FC
a
b
b
c
d
a
DT 0h
DT 72 h
DT 96 h
DT 120 h
DT 144 h
DT 168 h
Treatment

Figure 2.7: Validation of the relative expression levels of five selected genes responsive to combined DTHT stress from RNA-Seq analysis by quantitative real-time PCR (qPCR). The genes selected were differentially expressed, and the time point at which these genes showed high expression from the RNA-Seq data were selected with its control for qPCR validation. **7b.** Validation of relative expression of DT-responsive gene UDP-glucosyl transferase 85A3. UDP-glucosyl transferase 85A3 was up-regulated and down-regulated at different time points during DT stress from the RNA-Seq data. The expression pattern of the qPCR analysis is like results from the RNA-Seq analysis. The different alphabets in the Figure show that the samples collected from the different time point of DT are significantly different from the control at p-value<0.05. qPCR results from two technical replicates and three biological replicates were analyzed using ANOVA from Minitab 18 software. The x-axis shows the treatment imposed on switchgrass. The y-axis shows the relative expression of the genes.

2.4 Discussion

DT or HT stress alone has been found to affect switchgrass physiology and cause a reduction in biomass yield [22,29]. Extensive reports on transcriptome changes in plants during DT stress have been reported in both plant models and crop species [40]. The transcriptional response of switchgrass when imposed with solely DT or HT stress has been reported in previous studies [22,29,30]. However, transcriptome data associated with switchgrass when imposed with the combination of DTHT are not available. Molecular mechanisms during DTHT in plants such as lentil, cereals, and Kentucky bluegrass [41–43] have been reported. The primary objective of this study was to understand the transcriptional changes and molecular mechanisms in switchgrass in response to DT and the combined effects of DTHT.

2.4.1 Genes differentially expressed due to solely DT stress

In this study, water deficit in switchgrass triggered an up-regulation of more genes than down-regulated genes (**Figure 2.2**). One of the DT-responsive genes

identified from our analysis (Pavir.9KG421700.v4.1) and reported in the drought is galactinol synthase (Gols1). Gols1 catalyzes the biosynthesis of raffinose family oligosaccharides (RFOs). The RFO biosynthetic pathway is a major metabolic activity in plants and has been found to respond to various abiotic stresses. RFOs have emanated as essential molecules in plants during stress due to their antioxidant and membrane stabilizing properties. RFOs can be found in the chloroplast, which indicates its role in regulating genes in the photosystem II pathway [45,45]. Among DT-responsive genes that were shown to be induced in our analysis is OST1 (Pavir.9KG103200.v4.1). OST1 is found in stomatal guard cells and is known to activate SLAC1 which is required for stomatal closure during DT in plants [46]. DT stress activates the production of the hormone ABA. Mustilli et al. (2002) reported ABA-induced stomatal closure, which is impaired in *ost1*[47].

AREB1 (Pavir.J643700.v4.1) was also identified as a DT-responsive gene from our analysis and in the droughtDB (**Table 2**). It has also been found that the AREB subfamily of proteins and orthologues of AREB are found to be involved in ABA signaled transduction [48]. ABA plays an important role in plants and is involved in various physiological and developmental processes, including stomatal closure and response to a myriad of abiotic stresses such as cold, DT, and salinity [49]. Targets of ABA-dependent pathways recruit transcription factors such as AREB at the promoter sites to activate transcription. During DT stress, the level of ABA increases, causing ABA receptors PYR/PYL/RCAR to recruit phosphatase PP2C (identified in the KEGG pathway analysis in **Tables 1 and 2**) for downstream activation in the ABA-dependent signaling pathway [50]. ABA is known to regulate a large number of dehydration-responsive genes, which is associated with DT tolerance. These genes are

not limited to late embryogenesis abundant (LEA), Responsive to ABA 18 (RAB18), and RD22. Apart from the ABA-dependent genes, other DT-responsive genes are also ABA-independent. An example of an ABA-independent gene belongs to the family of dehydration-responsive element-binding (DREB) protein. DREB2 was up-regulated in the switchgrass plants imposed with DT. In various studies, DREB is more involved in DT stress and has been identified in rice and maize [30]. As expected, LEA, RD22, and RAB18 were induced with DT stress from our study. There were 35 DT responsive genes and 27 DTHT responsive genes with 12 overlapping genes in the droughtDB. Some of the genes identified as DT-responsive from our study have been listed in the manually curated compilation of molecularly characterized genes that are involved in DT stress response (**Tables 2 and 3**). These genes include AREB/ABF and glutathione S-transferases (GSTs). Previous reports indicates that overexpression of ABF4/AREB2 lead to ABA-hypersensitive phenotypes in Arabidopsis. Similarly, transgenic Arabidopsis plants with enhanced AREB/ABF expression showed enhancement in DT tolerance, indicating the role of AREB/ABF in ABA response and stress tolerance [48,51]. GSTs have been reported to a significant role in oxidative stress metabolism. Glutathione S-transferase U17A (GSTU17) is among the genes identified in the switchgrass samples under DT stress. In another study, mutants of GSTU17 in Arabidopsis became more tolerant to DT stress and salt stress than wild-type plants suggesting the role of GSTU17 in DT and salt stress tolerance [52].

Photosynthesis is among the processes affected by plant dehydration. In response to the water deficit in the switchgrass plants, transcripts encoding Rubisco activase, Rubisco methyltransferase family protein, photosystem II subunit O-2 (PSII), phosphoenolpyruvate carboxylase family protein initiation of CO_2 into oxaloacetate in

C4 plants [53] and phosphoenolpyruvate carboxylase: encoded by Ppc genes for initial fixation of CO2 were down-regulated. Two genes, carbonic anhydrase (associated with carbon-fixing and metabolism in C4 plants) and phosphoenolpyruvate carboxykinase 1 which was previously identified by Ayyappan et al. (2017) as a C4 photosynthetic enzyme were downregulated in response to the DT stress [54]. These findings are consistent with a report on the down-regulation of genes associated with photosynthesis during abiotic stress. Interestingly, we saw in our analysis that another transcript, Pavir.4NG244100.v4.1annotated as photosystem II subunit P-1 was down-regulated. Down-regulation of PSII affects electron transport, leading to the generation of harmful reactive oxygen species (ROS). A controlled amount of ROS protects the plant from DT as part of the signaling (ABA-dependent) pathways. However, an excessive amount of ROS which can be produced due to prolong DT could destroy critical cellular machinery of the plant while under DT stress [55]. From our analysis, Pavir.6NG292200.v4.1 annotated as Fe superoxide dismutase 3, and Pavir.3KG389500.v4.1, annotated as manganese superoxide dismutase 1 were up-regulated as scavengers of ROS to enhance the antioxidant defense of the plants under DT stress. In a previous study, the expression of Mn-SOD in transgenic *Medicago sativa* (alfalfa) plants showed increased tolerance against DT injury.

Similarly, alfalfa's in cold conditions showed an increased expression of Mn-SOD and Fe-SOD [56,57]. Understanding the antioxidant defense pathway will help to enhance switchgrass under DT stress. It is interesting to note that from our analysis Pavir.1KG123700.v4.1 annotated as 3-ketoacyl-CoA synthase 11 was up-regulated at four different time points of DT conditions. A recent study shows that 3-ketoacyl-CoA synthase (involved in lignin biosynthesis) could help to improve DT tolerance in tea

plants [58]. Similarly, Pavir.9NG554400.v4.1 annotated as basic helix-loop-helix (bHLH) DNA-binding superfamily protein was down-regulated at four different time points of DT. Waseem et al. (2019) showed that overexpression of bHLH enhanced abiotic stress tolerance in tomatoes [59]. These genes could provide insight in providing DT tolerance in switchgrass especially during prolonged exposure to DT.

KEGG pathway enrichment results showed that twelve genes were enriched in the term glutaredoxins. Glutaredoxins have been shown to be involved in different stress responses and regulation of the Krebs cycle and signaling pathways. Overexpression of some members of the glutaredoxin family modulated plant response to various stresses. For example, transgenic tomato plants with overexpression of SIGRX1 exhibited tolerance to hydrogen peroxide, DT, and salt stress [60]. One of the significant pathways enriched by the DT-responsive genes from this report was response to water. Another report by Bhardwaj et al. (2015) identified GO terms for DT *Brassica juncea* samples which include response to water deprivation (GO:0009414) [61].

2.4.2 Genes differentially expressed due to DTHT stress

From our analysis, most of the genes in response to combined DTHT were down-regulated (**Figure 2.2**). A combination of DTHT stress in Arabidopsis caused up-regulation of more transcripts compared to down-regulated transcripts, although this is in contrast to our findings [15]. In another report, several abiotic stress factors not limited to DT and HT stress led to down-regulation of multiple genes, indicating general transcriptional repression [62]. The transcriptome responses of the control switchgrass plant and those subjected to individual DT and combination of DTHT stress were different. However, there were common DEGs in response to DT stress

and a combination of DTHT stress. A significant overlap of transcripts expressed in DT or HT stress and combination of DTHT was found in plants in response to cold, DT, HT, and salt stress [11,15]. A similar finding was observed in tomato cultivars exposed to individual DT stress and combined DTHT. Single DT treatment on tomato cultivars had a considerable effect on HT stress [63]. This finding could explain why more genes responsive to DT were identified in combined DTHT stress plants. Jia et al. (2017) [64] identified an overlap of genes such as those involved in hormone metabolism (ABA) in *Populus simonii* when a single DT or HT was compared to combined DT and HT stress. The overlap suggests specific defense mechanisms by plants in response to abiotic stresses, which can be further explored. We identified 35 DT and 27 DTHT responsive genes in switchgrass, of which 12 were common between the two conditions. The key genes that played an important role in switchgrass performance under DT and DTHT include RFO, OST1, AREB1, GSTU17. Open Stomata 1 (OST1) is involved in the ABA regulation of stomatal response ([65]. RFO is a biosynthetic pathway, and it's involved in a major metabolic activity in plants and has been found to respond to various abiotic stresses [44] . AREB1 is a transcriptional activator, and it controls the ABA signaling to improve DT tolerance [66]. Documentation of the response of GSTs to a plethora of environmental stress responses has also been documented. GSTU17 in Arabidopsis was seen to provide DT and salt stress tolerance [67,68]. This finding suggests the possible expression of GSTU17 in both DT and DTHT samples. Most of the genes were revealed in the droughtDB (Table 2 and 3).

In response to both DTHT, factors such as LEA and HT shock proteins (HSPs) were up-regulated in our analysis. LEA and HSPs have been reported as responsive to

DT and extreme temperatures, and they play an essential role in protecting the plant during stress. Wang et al. (2003) [69] reported the response of LEA and HSPs to DT, salinity, and HT stress. Interestingly, Pavir.5KG018400.v4.1 (LEA14) was significantly up-regulated at 168 h. The same transcript was up-regulated at time point 168/96 h in both DT and HT-treated samples. LEA proteins accumulate primarily in plants during water deprivation. However, LEA proteins have been reported to respond to extreme temperatures as well. A previous report in *Brassica juncea* indicated that LEA showed a 40-fold increase during DT stress and a 10-fold increase in HT stress [61]. This finding suggests that LEA14 could be a candidate gene for breeding in areas with severe DT and extreme temperatures.

We identified several HT shock proteins (HSPs) in the switchgrass samples imposed with DTHT stress. Pavir.9NG640000.v4.1 and Pavir.9KG490200.v4.1 transcripts annotated as HT-shock protein 70T-2 and HT shock protein-70 respectively were significantly up-regulated at four different time points of the study. Other HSPs identified include Pavir.1NG519200.v4.1 (HSP20-like chaperones superfamily protein), Pavir.1KG194500.v4.1 (HT shock protein 17.6A), Pavir.9NG570500.v4.1 (HT shock protein 21), Pavir.6KG320100.v4.1 (Chaperone protein htpG family protein), Pavir.9KG212600.v4.1 (HT shock protein 60). In a previous study, Grigorova et al. (2011) observed the induction of HSPs in wheat samples imposed with DTHT stress compared to single DT stress [16].

Additionally, Pavir.9KG480900.v4.1 annotated as ascorbate peroxidase 1 (APX) and Pavir.7KG159800.v4.1 (stromal ascorbate peroxidase) were also found to be up-regulated by our analysis. The role of the *APX* gene in response to abiotic stress

conditions such as temperature, high light, DT, salinity, and heavy metals has been reported [70].

The Pavir.9NG211300v4.1 transcript encoding the ABO1/ELO2 gene was identified in the DT database and only responsive to DTHT stress. ABO1/ELO2 is an ABA-induced gene, and mutants showed affected development of guard cells, causing a decrease in the number of stomatal cells. ABO1/ELO2 is a subunit of Elongator, a multifunctional complex with roles in transcription which provided an uncommon mechanism of DT tolerance in Arabidopsis [71]. From our analysis, ABO1/ELO2 was up-regulated and this could be in response to the combined effect of DTHT to induce ABA hormones to regulate the stomata cells. Interestingly, another transcript Pavir.2KG247300.v4.1 codes for poly(ADP-ribose) polymerase (PARP1) were responsive in only DT and HT-treated switchgrass samples. PARP regulates transcription, metabolism and is involved in organizing the chromatin structure. Also, PARP responds to both biotic and abiotic stresses. From our analysis, PARP was up-regulated in response to DTHT stress. In a previous study, down-regulation of PARP1 increased DT tolerance in Arabidopsis [72]. This suggests that up-regulation of PARP1 in response to DTHT in the switchgrass samples could reduce its DT tolerance.

2.4.3 Genes deduced as HT responsive genes

As our primary focus in this experiment was on DT and DTHT responsive genes, we did not include HT only treatment. However, when we analyzed DTHT vs DT data for probable HT responsive genes, we found some interesting results. The HT responsive genes, i.e., HSPs that we detected are similar to HT genes found in wheat and switchgrass when exposed to only HT stress [16,22]. The HT responsive

genes identified in this experiment could serve as basis for future studies when imposing only HT stress.

2.4.4 TFs responsive to individual DT and DTHT stress

The differential expression pattern of DT-responsive genes was accompanied by different families of TFs, including bHLH (basic helix-loop-helix), WRKY, NAC (NAM, ATAF and CUC) and ERF (ETHYLENE RESPONSE FACTOR). Transcription factors known to be involved in DT stress response include WRKY, C2H2 and NAC, and these were more abundant in DT compared to DTHT samples (as shown in the TF statistics in **Additional File 10**). This finding may suggest that these TFs were induced early to initiate a transcriptional response to DT stress. Interestingly, the TFs mentioned above were identified in Populus species (*Populus davidiana*) under DT stress [73]. The bHLH TF was identified to be more highly expressed in response to DT stress alone, compared to DTHT stress in switchgrass. Mun et al. identified a strong expression pattern of bHLH in *P. davidiana* at 6 h and 12 h time points of their study [73]. Also, PebHLH35 as one of the families of bHLH, has been recognized to play a significant role in DT tolerance by controlling stomatal development and photosynthesis in Arabidopsis [74]. TFs such as MYB, bHLH, and WRKY were also abundantly identified in *Brassica juncea* plants under DT stress [61]. A high number of MYB and CH3 TFs were identified in DTHT samples compared to DT samples. MYB TF is known to control various processes including development, metabolism, and responses to biotic and abiotic stresses. A previous report showed that AtMYB096 from Arabidopsis is associated with ABA and JA-mediated pathway and provided DT tolerance in Arabidopsis. In another study, BcMYB1 TF from *Boea crassifolia* is reported to provide DT tolerance [75]. There

were relatively more NAC related TFs identified in response to DTHT stress (**Additional file 10**). However, some NAC TFs were either down-regulated or up-regulated, a differential expression of the TFs have been indicated in **Additional file 9** For example, NAC domain-containing protein 47, NAC domain-containing protein 83, and NAC domain-containing protein 41 were down-regulated whereas NAC domain-containing protein 102 and NAC domain-containing protein were up-regulated. Various NAC genes have been studied in switchgrass. An example is an identification and functional characterization of PvSWNs in switchgrass. These NAC genes have been reported to be associated with lignin and biosynthetic pathway [76]. Various ERF (ethylene-responsive factor family) TFs were responsive to single DT stress and DTHT stress from our analysis (**Table 1**). ERF TF family has been characterized in a previous study, and they have been found to respond to HT stress in *Populus simonii* [64]. Similarly, ERF isolated from soybean (*GmERF7*) was induced by DT and salt stress. However, *GmERF7* was reported to be down-regulated during cold stress in the same study by Zhai et al. (2013) [77]. In both DT and DTHT responsive TFs, bHLH TFs had the highest number. In a previous study, bHLH TFs have been reported to be related to DT [74,78,79]. Other stress-responsive TF families such as WRKY, MYB, and NAC previously reported were identified [80]. After bHLH, the next highest TF family identified from the analysis is NAC (NAM, ATAF1,2, and CUC2). NAC is one of the largest TFs and has been shown as an important regulator of abiotic stresses [81,82]. Reports indicate that NAC regulates DT stress when overexpressed in plants. Similarly, NAC genes, when overexpressed in Arabidopsis (*ANAC019*, *ANAC055,* and *ANAC072*) and rice (*OsNAC5, OsNAC6, OsNAC10*) enhanced DT and salt tolerance [83–85]. We also identified a high amount of bZIP TF encoding genes in both DT and

DTHT samples. Similar to bHLH and NAC, bZIP TF family has been reported to respond to various abiotic stresses. In rice, bZIP has been related to DT with OsbZIP16 being listed as a key candidate gene for DT tolerance [86]. Interestingly, more C3H TF was induced during DTHT stress compared to only DT stress. Our study reveals C3H as a candidate TF for both DTHT tolerance studies in plants. Analysis of C3H TF family in *Aegilops tauschii* suggested that overexpression of *AetTZF1* caused the plant to be more tolerant to DT stress [87].

2.4.5 Effect of DT and HT stress on phenylpropanoid metabolism

Phenylpropanoid is associated with lignin or flavonoid biosynthesis and plays essential role in the production of quality feedstock. Although phenylpropanoid pathway was not identified from the KEGG pathway or GO analysis, genes that are involved in the phenylpropanoid pathway previously identified by Ayyappan et al. (2017) [88] such as cinnamate-4-hydroxylase (C4H) with gene ID Pavir.J661300.v4.1, hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase (HCT) (gene ID Pavir.6KG280500.v4.1) were down-regulated with an extreme temperature at time point 168/96 h. Except for cinnamyl alcohol dehydrogenase 9 (CAD9) (Pavir.7NG065100.v4.1), which was up-regulated (**Additional file 5**). The role of CAD9 in lignin composition have been reported by Kim et al. (2004) [89]. CAD9 has been reported to catalyze the final step required to complete the production of lignin monomers such as coniferyl alcohol, sinapyl alcohol, and 4-coumaryl alcohol [90]. The presence of lignin limits the bioconversion of carbohydrates to ethanol from switchgrass. This limitation can lead to the high cost of cellulosic ethanol production; therefore, an effective approach previously reported was to cause downregulation of the genes involved in lignin biosynthesis to reduce lignin production [91,92]. From

our analysis, CAD9 was found to be up-regulated in the DT and HT-treated samples. This finding suggests that DT and HT stress could cause an increase in lignin synthesis. Lignin biosynthesis negatively correlates with biomass and bioenergy production in switchgrass because of the recalcitrant nature of the cell wall [93]. In another study, down-regulation of the CAD gene in switchgrass by RNA silencing led to a reduction in the amount of lignin and increased biomass production [76]. We observed down-regulation of phenylpropanoid genes, HCT, and C4H. Down-regulation of HCT and C4H could be due to the general down-regulation of genes involved in metabolism in response to stresses. These genes can serve as a target for genetic manipulation to produce quality biomass in switchgrass.

In addition to regulating development, differentiation, metabolism, biotic and abiotic processes, TFs belonging to MYB proteins have been found to play a significant role in phenylpropanoid metabolism [75]. From our analysis, several MYB TFs were responsive to DTHT compared to the individual DT stress. The transcript Pavir.6KG070500.v4.1 which is annotated as a MYB-related family protein, was significantly down-regulated at three different time points from the analysis. MYB proteins also serve to regulate other branches of phenylpropanoid metabolism. TF AmMYB305 from *Antirrhinum majus*, and MYB from Arabidopsis have been identified with a function in phenylpropanoid metabolism [94,95]. Switchgrass R2R3-MYB (PvMYB4) TF has been identified and characterized. PvMYB4 is reported to bind to AC-I, AC-II and AC-III elements of the monolignol pathway causing down-regulation of the genes in vivo. PvMYB4 is known to suppress phenylpropanoid metabolism and the quantity of lignin in switchgrass and tobacco. Overexpression of PvMYB4 caused a reduction in the lignin content and decreased recalcitrance in

transgenic switchgrass [96]. Hence, down-regulation of MYB related proteins from our analysis during DTHT stress may increase lignin production to affect biomass and biofuel production in switchgrass. This finding suggests that the MYB transcription factor should be considered in enhancing biomass under DT and extreme temperature conditions.

2.4.6 Validation of differentially regulated genes

We selected seven genes from the list of significantly regulated genes to validate experimentally by performing RT-PCR and qPCR. Five of the selected transcripts were either down or up-regulated in response to combined DT and HT stress. These transcripts include Pavir.3KG247300.v4.1, Pavir.9KG154500.v4.1, Pavir.9KG545000.v4.1, Pavir.4KG077400.v4.1, and Pavir.4KG264600.v4, which were annotated as a copper amine oxide, ATP dependent protease, UB-like protease 1A, leucine-rich receptor-like protein, and phosphatidylethanolamine-binding protein respectively. Copper amine oxide and UB-like Protease 1A were up-regulated in response to DT and HT stress while ATP-dependent protease, the leucine-rich receptor-like protein, was down-regulated response to DTHT stress. Another transcript Pavir.6KG130600.v4.1 which is annotated as UDP-glucosyl transferase 85A3 was up-regulated and down-regulated at different time points in response to single DT stress as indicated in Figure 2.7b. UDP-glucosyltransferase 85A3 from switchgrass was down-regulated with severe DT at DT-168 h. A UDP-glycosyltransferase 76C2 (UGT76C2) belonging to the same family as UGT85A played a significant role in response to water deficit in a previous report Arabidopsis. Like our finding UGT76C2 from Arabidopsis was down-regulated in response to DT stress [97].

Our analysis found that the transcript Pavir.9NG755000.v4.1 which is annotated as basic chitinase, was only identified in samples exposed to DT and HT switchgrass samples. This gene was down-regulated in all the time points but was significantly up-regulated at extreme DT and HT (**Additional file 5**). RT-PCR confirmed results from the RNA-Seq data, and which is consistent with the previous report on the function of chitinase genes (figure not shown). Chitinase enzymes are reported as defense proteins and their expression are usually influenced by environmental stress [98]. They provide resistance against pathogens and is tolerant to various environmental stresses. Chitinase genes have been recognized to respond to environmental stresses. In a previous study, the expression of one of the chitinase enzymes was enhanced in Arabidopsis samples with allosamidin and strong HT stress compared to control plants [99]. Similar to our findings, Pavir.9NG755000.v4.1 annotated as chitinase may have been differentially expressed due to the HT stress. The up-regulation of the chitinase gene may help to improve DT and HT stress tolerance in switchgrass.

2.5 Conclusion and future perspectives

Several studies have been conducted in switchgrass in response to individual biotic or abiotic stress. However, scientific information on the transcriptional changes in switchgrass under combined DT and HT stress is underexplored. We utilized RNA-Seq approaches to elucidate transcriptomic changes in switchgrass when exposed to either DT or a combination of DT and HT. Many of the genes identified were in response to DTHT stress. Additionally, we identified TFs that were regulated by these stresses. We found an overlap of genes in response to a single DT and a combination

of DTHT stress. Interestingly, these transcripts were found in the droughtDB. Both single DT and DTHT had an effect on the photosynthetic machinery and produced genes involved in oxidative stress damage which can affect biomass production. Several HSPs and chaperones were produced in the combined DT and HT switchgrass samples compared to those with individual DT stress. The GO annotation and KEGG pathway analysis showed connections between the identified GO terms. Genes associated with the photosynthesis machinery and control carbon fixation were down-regulated, suggesting the effect of DTHT on biomass production. A co-expression analysis revealed a unique expression pattern of the differentially expressed genes, which were classified into modules. Moreover, the significant pathways enriched in most of the DEG genes were involved in the metabolic and ABA signaling pathways.

Further, the combined DT and HT stress resulted in a unique regulation of genes and TFs involved in the phenylpropanoid pathways such as CAD9, C4H and HCT. CAD9, C4H and HCT are associated with lignin biosynthesis, which negatively correlates with biomass and bioenergy production. The stress-responsive genes and TFs identified in this study will be helpful in developing switchgrass cultivars with improved tolerance to DT and HT stress. The transcriptome data generated in this study could be used as a reference to investigate further DT and HT stress tolerance in bioenergy crops and plants in general.

2.6 Materials and methods

2.6.1 Growth and treatment of plants

The experiments were conducted using a lowland ecotype Alamo, AP13 genotype. The AP13 genotype was a selection from the publicly available switchgrass cultivar 'Alamo'. Initial selection was made at the University of Georgia, but later the genotype was moved to the greenhouse at Noble Research Institute, LLC. Clonal copies of the genotypes have been maintained in Noble greenhouse. Ramets of AP13 were transplanted into 3GP nursery pots (Growers Solution, Cookeville, TN) and grown for 40 days under optimum growing condition in the greenhouse and transferred to growth chambers at the Noble Research Institute, Ardmore, OK. The experiment was designed to mimic conditions in the natural environment where plants experience more than one type of stress. The goal is to identify the unique response of switchgrass to combined DTHT stress. The experiment was started five days after transfer to the growth chamber. The experiment was laid out in a randomized complete block design with three biological replicates. Six pots were assigned to control, 9 pots to DT, and 9 pots to DT and HT treatments at random during the transfer. The pots assigned to the three treatments were divided into three groups and assigned to the three replicates at random. The control and DT treatments were transferred to a growth chamber and the DT imposed with HT treatment was arranged in another growth chamber at random (Figure 2.8). Leaf tissue samples were collected as indicated in **Figure 2.8** at the same time (starting at 2:00 PM) of the day for all samples collected. Plant tissues were immediately frozen in liquid nitrogen and stored at -80 °C. The samples were then shipped to Delaware State University on dry ice overnight. Soil moisture at 10 cm depth of the pot was measured concomitant with

tissue sample collection using Field Scout TDR 100 Soil Moisture Meter (Spectrum Technologies, Aurora, IL). Leaf SPAD reading was also taken at the same time. A diagram to indicate how the growth chamber was separated for DT and HT treatments have been shown in **Figure 2.8.**

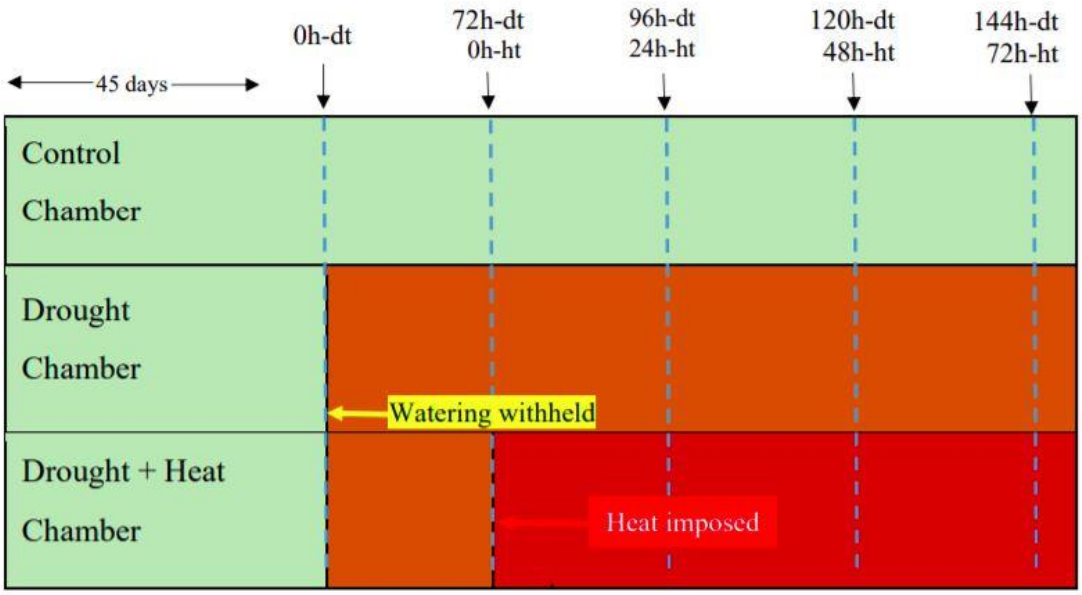


Figure 2.8: Control chamber: Regular watering (80% FC) and optimum temperature (30°/23°C day/night temperature); DT chamber: withhold watering at 45 days after transplanting the ramets and kept at optimum temperature (30°/23°C day/night temperature); DT + HT chamber: imposed HT after 72h of DT (35°/25°C day/night temperature); Leaf tissue samples were collected at 0h-DT (dt), 72h-dt/0h-HT (ht), 96h-dt/24h-ht, 120h-dt, 48h-ht, and 144h-dt/72h-ht impositions.

2.6.2 RNA isolation and cDNA synthesis

Total RNA was extracted from leaves of control, DT, HT, and combined DT and HT-treated switchgrass using RNeasy Plant mini kit (Qiagen Inc., CA) according to the manufacturer's instruction. To eliminate contaminating genomic DNA, all RNA

samples were treated with amplification grade DNase I (Invitrogen) following manufacturer's protocol. The concentration and purity of the RNA samples were determined using Nanodrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE). The A260/A280 nm ratios for a majority of the samples were 2.1. The quality of the RNA samples was determined by 1% agarose gel electrophoresis and Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA) for 28S/18S rRNA band intensity (2:1) and RNA integrity number (RIN) >8. The RNA samples were stored at -80 ^{o}C for use in downstream experiments. 1 µg of DNase treated RNA was used for cDNA synthesis using Protoscript II First Strand cDNA Synthesis kit (New England Biolabs, Ipswich MA) following the manufacturer's instructions. In synthesizing the complementary DNA, 1 µg of DNase treated RNA was denatured with Oligo dT at 65 ^{O}C for 5 min; followed by adding Protscript II reaction mix and Protoscript II enzyme mix which were incubated at 42 ^{O}C for 60 mins. The Protoscript II enzyme was denatured at 80 ^{O}C for 5 mins and the cDNA was then stored at -20 ^{O}C.

2.6.3 Library construction and sequencing

A Fragment Analyzer (Advanced Analytical, Ames, IA) was used to check the quality and purity of all the RNA samples. RNA-Seq libraries were prepared using Illumina TruSeq Stranded mRNA Sample Preparation Kit (Illumina Inc., San Diego, CA) following the manufacturer's instructions at the Delaware Biotechnology Institute, Newark, DE, USA. The libraries were sequenced on Illumina HiSeq 2500 platform with 101 bp paired-end reads.

2.6.4 Processing of RNA-Seq data

FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/; v0.0.14) was used to perform quality control for RNA-Seq data requiring at least a 30-base quality score and at least 50 bps of read length. TopHat (v2.1.1) [100] was then used to align the reads to the switchgrass reference genome (**Additonal file 1**) . FPKM values were calculated using the Cufflinks (v2.2.1) suite of tools [101]. To get the read count for the genes, HTSeq (v0.7.0) was used [102].

2.6.5 Filtration of genes based on FPKM values

Low-expressed features tend to reflect noise and correlations based on counts that are mostly zero and are not meaningful. Based on the annotation file released, there are 91,838 genes across the switchgrass genome. FPKM values for each gene in each sample were calculated using cuffnorm in the Cufflinks suite of tools [101]. A given gene is retained for further analysis if at least half of the 15 groups have average FPKM value >1 and the average FPKM value of all samples included is >1 [103]. In total, 32,190 genes were retained for downstream analysis.

2.6.6 Identification of DT and HT responsive genes

To identify DT responsive genes in the RNA-Seq samples, DESeq2 package was used [104]. First, genes that were differentially expressed between DT treatment group and control group at 0 h were excluded. Then the remaining genes that were differentially expressed between DT treatment group and control group in at least one of the following time points (72, 96, 120, 144 or 168 h) were defined as DT responsive genes. To identify responsive genes related to combination of DT and HT (DTHT), genes that were differentially expressed between group with combination of DT and HT treatment and control group at 0 h and 72 h were excluded. Then the remaining

genes that were differentially expressed between group with combination of DT and HT treatment and control group in at least one of the following time points (96, 120, 144 or168 h) were defined as DTHT responsive genes. Although the switchgrass plants were not exposed to direct heat temperatures separately, an assumption was made that the DEGs in the combined DTHT vs DT samples could be due to the heat stress imposed. To identify responsive genes that may be related to HT stress, genes that were differentially expressed between group with combination of DTHT treatment and DT group at 0 h and 72 h were excluded. Then the remaining genes that were differentially expressed between group with combination of DT and HT treatment and DT treatment group in at least one of the following time points (96, 120, 144 or 168 h) were defined as HT responsive genes.

2.6.7 Construction of co-expression network using WGCNA

Log2 transformed FPKM matrix of the genes (32,190) was used as input to WGCNA (v1.51) (**Additional file 4**). The function “pickSoftThreshold” was used to pick an approximate power value. Then “blockwiseModules” (networkType = "signed hybrid") was used to construct co-expression network.

2.6.8 Functional analysis of stress responsive genes

2.6.8.1 *GO enrichment analysis***:**

For stress responsive genes or the genes in the co-expression networks, the corresponding GO terms of the genes were extracted. Singular Enrichment Analysis (SEA) from agriGO [105] was used to perform GO enrichment analysis.

2.6.8.2 *KEGG enrichment analysis*:

For stress responsive genes, the corresponding KEGG orthology terms of the genes were also extracted. ClusterProfiler (v 3.0.5) [106] were then used to perform KEGG enrichment analysis.

2.6.8.3 *MapMan analysis:*

To further understand the biological functions of the DEGs and specific pathways or genes associated with single DT and combined DTHT samples, we conducted metabolic pathways analysis using the MapMan software (http://MapMan.gabipd.org). Default settings in MapMan software do not support mapping for the switchgrass genome. A customized input file was created using the Mercator [107] tool and protein sequences from switchgrass v4.1. The Mercator is a tool to batch classify protein or gene sequences into MapMan functional plant categories and create a draft metabolic network which can be directly used in MapMan software. Mercator output was used as mapping file for MapMan.

2.6.8.4 *Annotation of transcription factor*:

Genome-wide identification of TF were performed using PlantTFDB 4.0 [108]. Proteins of primary transcript for the genes were uploaded to the prediction server of PlantTFDB 4.0. The output of the prediction severs included TF types and best hits in Arabidopsis.

2.6.9 Quantitative real-time (qRT-PCR) analysis

QRT-PCR was performed using the synthesized cDNA. The primers were designed based on the differentially expressed transcripts of DT and combined DT and HT stresses (DTHT). These primers will be used to validate the quantitative

expression of the genes with highly expressed transcripts (log2FC > 2) from DTHT analysis. The selected DTHT and DT genes and the list of specific primer sequences are given in (**Additional file 14)**). The primers were designed using the online tool for real-time PCR (TaqMan) primer design by GenScript Inc. (Piscataway, NJ). A conventional PCR was first performed to validate the primers before using them in qRT-PCR. 1 µl of 50 ng of cDNA was used a template for the conventional PCR reaction under these conditions (95 ^{o}C for 1 min, 55 ^{o}C for 30 s and 72 ^{o}C for 1 min) for 35 cycles. The PCR product was separated on a 1% agarose gel stained with ethidium bromide.

qRT-PCR was performed using an ABI 7500 real-time PCR system and SYBR Green Kit (Applied Biosystems, Grand Island, USA). Twenty-five µLs of the PCR reactions containing1 µg of 1st-strand cDNA, 12.5 µL of Power SYBR Green Master Mix, and 3 µL of 10 nM specific primers (forward and reverse) and 9.5 µL of water. The reference gene *Actin11* was used as an internal control primer to normalize the results in all the samples. The PCR conditions for the qRT-PCR were the following; 95 ^{o}C for 10 min, followed by 40 cycles of 95 ^{o}C for 15 s and 65 ^{o}C for 1 min. The efficiency of the primers was tested, and the relative expression was determined from three biological and three technical replicates using ΔΔCT method (Schmittgen and Livak, 2010). Minitab-17 software (State College, PA) was used to analyze the normalized CT values from the qRT-PCR analysis.

2.7 Abbreviation

DT: Drought
DTHT: Drought and heat stress
HT: Heat

DEG: differentially expressed genes

QRT-PCR: quantitative real-time PCR

GO: Gene Ontology

KEGG: Kyoto Encyclopedia of Genes and Genome

ABA: Abscisic acid

2.8 Availability of data and materials

The datasets supporting the conclusions of this research article have been included in the article and as additional files. The sequencing database for switchgrass under DT and HT stress has been deposited at NCBI under GEO accession number (**GSE174278**) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174278 and it can be downloaded.

2.8.1 Supplementary information

The supplementary information can be accessed using this link (https://drive.google.com/drive/folders/1NrOnRj1DwLGtmVsSCatwpXw13zEwGdUp?usp=sharing)

REFERENCES

[1] Bai Y, Kissoudis C, Yan Z, Visser RGF, van der Linden G. Plant behaviour under combined stress: tomato responses to combined salinity and pathogen stress. Plant J 2018;93:781–93. https://doi.org/10.1111/tpj.13800.

[2] Sewelam N, Oshima Y, Mitsuda N, Ohme-Takagi M. A step towards understanding plant responses to multiple environmental stresses: A genome-wide study. Plant, Cell Environ 2014. https://doi.org/10.1111/pce.12274.

[3] Kissoudis C, Sunarti S, Van De Wiel C, Visser RGF, Van Der Linden CG, Bai Y. Responses to combined abiotic and biotic stress in tomato are governed by stress intensity and resistance mechanism. J Exp Bot 2016;67:5119–32. https://doi.org/10.1093/jxb/erw285.

[4] Rasmussen S, Barah P, Suarez-Rodriguez MC, Bressendorff S, Friis P, Costantino P, et al. Transcriptome Responses to Combinations of Stresses in Arabidopsis. Plant Physiol 2013;161:1783–94. https://doi.org/10.1104/pp.112.210773.

[5] Pandey P, Irulappan V, Bagavathiannan M V., Senthil-Kumar M. Impact of Combined Abiotic and Biotic Stresses on Plant Growth and Avenues for Crop Improvement by Exploiting Physio-morphological Traits. Front Plant Sci 2017;8:1–15. https://doi.org/10.3389/fpls.2017.00537.

[6] O'Rourke JA, McCabe CE, Graham MA. Dynamic gene expression changes in response to micronutrient, macronutrient, and multiple stress exposures in soybean. Funct Integr Genomics 2020. https://doi.org/10.1007/s10142-019-00709-9.

[7] Mittler R. Abiotic stress, the field environment and stress combination. Trends Plant Sci 2006. https://doi.org/10.1016/j.tplants.2005.11.002.

[8] Prasad PVV, Pisipati SR, Momčilović I, Ristic Z. Independent and Combined Effects of High Temperature and Drought Stress During Grain Filling on Plant Yield and Chloroplast EF-Tu Expression in Spring Wheat. J Agron Crop Sci 2011. https://doi.org/10.1111/j.1439-037X.2011.00477.x.

[9] Xiong L, Ishitani M, Zhu J-K. Interaction of Osmotic Stress, Temperature, and Abscisic Acid in the Regulation of Gene Expression in Arabidopsis. Plant Physiol 1999;119:205–12. https://doi.org/10.1104/pp.119.1.205.

[10] Anderson JP. Antagonistic Interaction between Abscisic Acid and Jasmonate-Ethylene Signaling Pathways Modulates Defense Gene Expression and Disease Resistance in Arabidopsis. Plant Cell Online 2004;16:3460–79. https://doi.org/10.1105/tpc.104.025833.

[11] Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, et al. Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J 2002;31:279–92. https://doi.org/10.1046/j.1365-313X.2002.01359.x.

[12] Yang M, Zhang Y, Zhang H, Wang H, Wei T, Che S, et al. Identification of MsHsp20 Gene Family in Malus sieversii and Functional Characterization of MsHsp16.9 in Heat Tolerance. Front Plant Sci 2017;8:1–17. https://doi.org/10.3389/fpls.2017.01761.

[13] Jin J, Yang L, Fan D, Liu X, Hao Q. Comparative transcriptome analysis uncovers different heat stress responses in heat-resistant and heat-sensitive jujube cultivars. PLoS One 2020. https://doi.org/10.1371/journal.pone.0235763.

[14] Rizhsky L, Liang H, Mittler R. The combined effect of drought stress and heat shock on gene expression in tobacco. Plant Physiol 2002;130:1143–51. https://doi.org/10.1104/pp.006858.

[15] Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, Mittler R. Rizhsky et al. 134 (4) 1683. (2004).pdf 2004;134:1683–96. https://doi.org/10.1104/pp.103.033431.1.

[16] Grigorova B, Vaseva I, Demirevska K, Feller U. Combined drought and heat

stress in wheat: Changes in some heat shock proteins. Biol Plant 2011;55:105–11. https://doi.org/10.1007/s10535-011-0014-x.

[17] Muthuramalingam P, Krishnan SR, Pothiraj R, Ramesh M. Global transcriptome analysis of combined abiotic stress signaling genes unravels key players in Oryza sativa L.: An in silico approach. Front Plant Sci 2017;8:1–13. https://doi.org/10.3389/fpls.2017.00759.

[18] Hivrale V, Zheng Y, Puli COR, Jagadeeswaran G, Gowdu K, Kakani VG, et al. Characterization of drought- and heat-responsive microRNAs in switchgrass. Plant Sci 2015;242:214–23. https://doi.org/10.1016/j.plantsci.2015.07.018.

[19] Sanderson MA, Adler PR, Boateng AA, Casler MD, Sarath G, Hitchcock A, et al. Switchgrass as a biofuels feedstock in the USA. Can J Plant Sci 2006;86:1315–25. https://doi.org/10.4141/P06-136.

[20] Adler PR, Sanderson MA, Boateng AA, Weimer PJ, Jung HJG. Biomass yield and biofuel quality of switchgrass harvested in fall or spring. Agron J 2006;98:1518–25. https://doi.org/10.2134/agronj2005.0351.

[21] Serba DD, Uppalapati SR, Krom N, Mukherjee S, Tang Y, Mysore KS, et al. Transcriptome analysis in switchgrass discloses ecotype difference in photosynthetic efficiency. BMC Genomics 2016;17:1–14. https://doi.org/10.1186/s12864-016-3377-8.

[22] Li YF, Wang Y, Tang Y, Kakani VG, Mahalingam R. Transcriptome analysis of heat stress response in switchgrass (Panicum virgatum L.). BMC Plant Biol 2013. https://doi.org/10.1186/1471-2229-13-153.

[23] Sun G, Stewart CN, Xiao P, Zhang B, McLaughlin S, Kiniry J, et al. MicroRNA Expression Analysis in the Cellulosic Biofuel Crop Switchgrass (Panicum virgatum) under Abiotic Stress. PLoS One 2012;7:e32017. https://doi.org/10.1371/journal.pone.0032017.rj

[24] Krasensky J, Jonak C. Europe PMC Funders Group Drought , salt , and temperature stress-induced metabolic rearrangements and regulatory networks Europe PMC Funders Author Manuscripts. J Exp Bot 2015;63:1593–608.

https://doi.org/10.1093/jxb/err460.Drought.

[25] Liu Y, Zhang X, Tran H, Shan L, Kim J, Childs K, et al. Assessment of drought tolerance of 49 switchgrass (Panicum virgatum) genotypes using physiological and morphological parameters. Biotechnol Biofuels 2015;8:1–18. https://doi.org/10.1186/s13068-015-0342-8.

[26] Jeandroz S, Lamotte O. Editorial: Plant Responses to Biotic and Abiotic Stresses: Lessons from Cell Signaling. Front Plant Sci 2017;8:1–3. https://doi.org/10.3389/fpls.2017.01772.

[27] van der Weijde T, Huxley LM, Hawkins S, Sembiring EH, Farrar K, Dolstra O, et al. Impact of drought stress on growth and quality of miscanthus for biofuel production. GCB Bioenergy 2017;9:770–82. https://doi.org/10.1111/gcbb.12382.

[28] Barney JN, Mann JJ, Kyser GB, Blumwald E, Van Deynze A, DiTomaso JM. Tolerance of switchgrass to extreme soil moisture stress: Ecological implications. Plant Sci 2009;177:724–32. https://doi.org/10.1016/j.plantsci.2009.09.003.

[29] Meyer E, Aspinwall MJ, Lowry DB, Palacio-Mejía J, Logan TL, Fay PA, et al. Integrating transcriptional, metabolomic, and physiological responses to drought stress and recovery in switchgrass (Panicum virgatum L.). BMC Genomics 2014;15:527. https://doi.org/10.1186/1471-2164-15-527.

[30] Aimar D, Calafat M, Andrade AM, Carassay L, Bouteau F, Abdala G, et al. Drought effects on the early development stages of Panicum virgatum L.: Cultivar differences. Biomass and Bioenergy 2014;66:49–59. https://doi.org/10.1016/j.biombioe.2014.03.004.

[31] Kandel TP, Wu Y, Kakani VG. Growth and Yield Responses of Switchgrass Ecotypes to Temperature. Am J Plant Sci 2013;4:1173–80. https://doi.org/10.4236/ajps.2013.46145.

[32] Xie_et_al-2014-Plant_Biotechnology_Journal.pdf-switchgrass salinity and drought.pdf n.d.

[33] Mundy J, Nielsen HB, Brodersen P. Crosstalk 2006;11:63–4. https://doi.org/10.1016/j.tplants.2005.12.008.

[34] Johnson MB, Kawasawa YI, Mason CE, Krsnik Ž, Coppola G, Bogdanović D, et al. Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. Neuron 2009. https://doi.org/10.1016/j.neuron.2009.03.027.

[35] Oldham MC, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, et al. Functional organization of the transcriptome in human brain. Nat Neurosci 2008. https://doi.org/10.1038/nn.2207.

[36] Barabási AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nat Rev Genet 2004. https://doi.org/10.1038/nrg1272.

[37] Alter S, Bader KC, Spannagl M, Wang Y, Bauer E, Schön CC, et al. DroughtDB: An expert-curated compilation of plant drought stress genes and their homologs in nine species. Database 2015;2015:1–7. https://doi.org/10.1093/database/bav046.

[38] Robert N, Merlot S, N'Guyen V, Boisson-Dernier A, Schroeder JI. A hypermorphic mutation in the protein phosphatase 2C HAB1 strongly affects ABA signaling in Arabidopsis. FEBS Lett 2006. https://doi.org/10.1016/j.febslet.2006.07.047.

[39] Kuromori T, Sugimoto E, Shinozaki K. Arabidopsis mutants of AtABCG22, an ABC transporter gene, increase water transpiration and drought susceptibility. Plant J 2011;67:885–94. https://doi.org/10.1111/j.1365-313X.2011.04641.x.

[40] Deyholos MK. Making the most of drought and salinity transcriptomics. Plant, Cell Environ 2010;33:648–54. https://doi.org/10.1111/j.1365-3040.2009.02092.x.

[41] Jiang Y, Huang B. Drought and heat stress injury to two cool-season turfgrasses in relation to antioxidant metabolism and lipid peroxidation. Crop Sci 2001;41:436–42. https://doi.org/10.2135/cropsci2001.412436x.

[42] Sehgal A, Sita K, Kumar J, Kumar S, Singh S, Siddique KHM, et al. Effects of

Drought, Heat and Their Interaction on the Growth, Yield and Photosynthetic Function of Lentil (Lens culinaris Medikus) Genotypes Varying in Heat and Drought Sensitivity. Front Plant Sci 2017;8. https://doi.org/10.3389/fpls.2017.01776.

[43] Barnabás B, Jäger K, Fehér A. The effect of drought and heat stress on reproductive processes in cereals. Plant, Cell Environ 2008;31:11–38. https://doi.org/10.1111/j.1365-3040.2007.01727.x.

[44] Knaupp M, Mishra KB, Nedbal L, Heyer AG. Evidence for a role of raffinose in stabilizing photosystem II during freeze-thaw cycles. Planta 2011;234:477–86. https://doi.org/10.1007/s00425-011-1413-0.

[45] Sengupta S, Mukherjee S, Basak P, Majumder AL. Significance of galactinol and raffinose family oligosaccharide synthesis in plants. Front Plant Sci 2015;6:1–11. https://doi.org/10.3389/fpls.2015.00656.

[46] Vahisalu T, Kollist H, Wang Y, Nishimura N, Chan W-Y, Valerio G, et al. SLAC1 is required for plant guard cell S-type anion channel. Natire 2008;452:487–91. https://doi.org/10.1038/nature06608.SLAC1.

[47] Mustilli A, Merlot S, Vavasseur A, Fenzi F, Giraudat J. Métodos para tomar decisiones. Plant Cell 2002;14:3089–99. https://doi.org/10.1105/tpc.007906.ABA.

[48] Furihata T, Maruyama K, Fujita Y, Umezawa T, Yoshida R, Shinozaki K, et al. Abscisic acid-dependent multisite phosphorylation regulates the activity of a transcription activator AREB1. Proc Natl Acad Sci U S A 2006;103:1988–93. https://doi.org/10.1073/pnas.0505667103.

[49] Gómez-Porras JL, Riaño-Pachón D, Dreyer I, Mayer JE, Mueller-Roeber B. Genome-wide analysis of ABA-responsive elements ABRE and CE3 reveals divergent patterns in Arabidopsis and rice. BMC Genomics 2007;8:1–13. https://doi.org/10.1186/1471-2164-8-260.

[50] Lourenço TF, Barros PM, Saibo NJM, Abreu IA, Santos AP, Antínio C, et al. Genomics of drought. Plant Genomics Clim. Chang., 2016.

https://doi.org/10.1007/978-1-4939-3536-9_5.

[51] Kang JY, Choi HI, Im MY, Kim SY. Arabidopsis basic leucine zipper proteins that mediate stress- responsive abscisic acid signaling. Plant Cell 2002;14:343–57. https://doi.org/10.1105/tpc.010362.tase.

[52] Chen J-H, Jiang H-W, Hsieh E-J, Chen H-Y, Chien C-T, Hsieh H-L, et al. Drought and Salt Stress Tolerance of an Arabidopsis Glutathione S-Transferase U17 Knockout Mutant Are Attributed to the Combined Effect of Glutathione and Abscisic Acid. Plant Physiol 2012;158:340–51. https://doi.org/10.1104/pp.111.181875.

[53] Westhoff P, Gowik U. Evolution of C4 phosphoenolpyruvate carboxylase. Genes and proteins: A case study with the genus Flaveria. Ann Bot 2004. https://doi.org/10.1093/aob/mch003.

[54] Ayyappan V, Saha MC, Thimmapuram J, Sripathi VR, Bhide KP, Fiedler E, et al. Comparative transcriptome profiling of upland (VS16) and lowland (AP13) ecotypes of switchgrass. Plant Cell Rep 2017. https://doi.org/10.1007/s00299-016-2065-0.

[55] Helena M, Carvalho C De. Drought stress and reactive oxygen species 2008;3:156–65.

[56] McKersie BD, Bowley SR, Harjanto E, Leprince O. Water-deficit tolerance and field performance of transgenic alfalfa overexpressing superoxide dismutase. Plant Physiol 1996. https://doi.org/10.1104/pp.111.4.1177.

[57] Samis K, Bowley S, McKersie B. Pyramiding Mn-superoxide dismutase transgenes to improve persistence and biomass production in alfalfa. J. Exp. Bot., 2002. https://doi.org/10.1093/jxb/53.372.1343.

[58] Gu H, Wang Y, Xie H, Qiu C, Zhang S, Xiao J, et al. Drought stress triggers proteomic changes involving lignin, flavonoids and fatty acids in tea plants. Sci Rep 2020. https://doi.org/10.1038/s41598-020-72596-1.

[59] Waseem M, Rong X, Li Z. Dissecting the role of a basic helix-loop-helix

transcription factor, SlBHLH22, under salt and drought stresses in transgenic Solanum lycopersicum L. Front Plant Sci 2019. https://doi.org/10.3389/fpls.2019.00734.

[60] El-Kereamy A, Bi Y-M, Mahmood K, Ranathunge K, Yaish MW, Nambara E, et al. Overexpression of the CC-type glutaredoxin, OsGRX6 affects hormone and nitrogen status in rice plants. Front Plant Sci 2015;6:1–12. https://doi.org/10.3389/fpls.2015.00934.

[61] Bhardwaj AR, Joshi G, Kukreja B, Malik V, Arora P, Pandey R, et al. Global insights into high temperature and drought stress regulated genes by RNA-Seq in economically important oilseed crop Brassica juncea. BMC Plant Biol 2015. https://doi.org/10.1186/s12870-014-0405-1.

[62] Weber C, Guigon G, Bouchier C, Frangeul L, Moreira S, Sismeiro O, et al. Stress by Heat Shock Induces Massive Down Regulation of Genes and Allows Differential Allelic Expression of the Gal / GalNAc Lectin in Entamoeba histolytica Stress by Heat Shock Induces Massive Down Regulation of Genes and Allows Differential Allelic Expr. Eukariotic Cell 2006;5:871–5. https://doi.org/10.1128/EC.5.5.871.

[63] Zhou R, Yu X, Ottosen CO, Rosenqvist E, Zhao L, Wang Y, et al. Drought stress had a predominant effect over heat stress on three tomato cultivars subjected to combined stress. BMC Plant Biol 2017;17:1–13. https://doi.org/10.1186/s12870-017-0974-x.

[64] Jia J, Zhou J, Shi W, Cao X, Luo J, Polle A, et al. Comparative transcriptomic analysis reveals the roles of overlapping heat-/drought-responsive genes in poplars exposed to high temperature and drought. Sci Rep 2017. https://doi.org/10.1038/srep43215.

[65] Acharya BR, Jeon BW, Zhang W, Assmann SM. Open Stomata 1 (OST1) is limiting in abscisic acid responses of Arabidopsis guard cells. New Phytol 2013. https://doi.org/10.1111/nph.12469.

[66] Fujita Y, Fujita M, Satoh R, Maruyama K, Parvez MM, Seki M, et al. AREB1

is a transcription activator of novel ABRE-dependent ABA signaling that enhances drought stress tolerance in Arabidopsis. Plant Cell 2005. https://doi.org/10.1105/tpc.105.035659.

[67] Kumar S, Trivedi PK. Glutathione S-transferases: Role in combating abiotic stresses including arsenic detoxification in plants. Front Plant Sci 2018. https://doi.org/10.3389/fpls.2018.00751.

[68] Wei L, Zhu Y, Liu R, Zhang A, Zhu M, Xu W, et al. Genome wide identification and comparative analysis of glutathione transferases (GST) family genes in Brassica napus. Sci Rep 2019. https://doi.org/10.1038/s41598-019-45744-5.

[69] Wang W, Vinocur B, Altman A. Plant responses to drought, salinity and extreme temperatures: Towards genetic engineering for stress tolerance. Planta 2003.https://doi.org/10.1007/s00425-003-1105-5.

[70] Caverzan A, Passaia G, Rosa SB, Ribeiro CW, Lazzarotto F, Margis-Pinheiro M. Plant responses to stresses: Role of ascorbate peroxidase in the antioxidant protection. Genet Mol Biol 2012;35:1011–9. https://doi.org/10.1590/S1415-47572012000600016.

[71] Chen Z, Zhang H, Jablonowski D, Zhou X, Ren X, Hong X, et al. Mutations in ABO1/ELO2, a Subunit of Holo-Elongator, Increase Abscisic Acid Sensitivity and Drought Tolerance in Arabidopsis thaliana. Mol Cell Biol 2006;26:6902–12. https://doi.org/10.1128/MCB.00433-06.

[72] De Block M, Verduyn C, De Brouwer D, Cornelissen M. Poly(ADP-ribose) polymerase in plants affects energy homeotasis, cell death and stress tolerance. Plant J 2005. https://doi.org/10.1111/j.1365-313X.2004.02277.x.

[73] Mun BG, Lee SU, Park EJ, Kim HH, Hussain A, Imran QM, et al. Analysis of transcription factors among differentially expressed genes induced by drought stress in Populus davidiana. 3 Biotech 2017. https://doi.org/10.1007/s13205-017-0858-7.

[74] Dong Y, Wang C, Han X, Tang S, Liu S, Xia X, et al. A novel bHLH

transcription factor PebHLH35 from Populus euphratica confers drought tolerance through regulating stomatal development, photosynthesis and growth in Arabidopsis. Biochem Biophys Res Commun 2014;450:453–8. https://doi.org/10.1016/j.bbrc.2014.05.139.

[75] Ambawat S, Sharma P, Yadav NR, Yadav RC. MYB transcription factor genes as regulators for plant responses: An overview. Physiol Mol Biol Plants 2013;19:307–21. https://doi.org/10.1007/s12298-013-0179-1.

[76] Zhong R, Yuan Y, Spiekerman JJ, Guley JT, Egbosiuba JC, Ye ZH. Functional characterization of NAC and MYB transcription factors involved in regulation of biomass production in switchgrass (Panicum virgatum). PLoS One 2015;10:1–24. https://doi.org/10.1371/journal.pone.0134611.

[77] Zhai Y, Wang Y, Li Y, Lei T, Yan F, Su L, et al. Isolation and molecular characterization of GmERF7, a soybean ethylene-response factor that increases salt stress tolerance in tobacco. Gene 2013. https://doi.org/10.1016/j.gene.2012.10.018.

[78] Castilhos G, Lazzarotto F, Spagnolo-Fonini L, Bodanese-Zanettini MH, Margis-Pinheiro M. Possible roles of basic helix-loop-helix transcription factors in adaptation to drought. Plant Sci 2014. https://doi.org/10.1016/j.plantsci.2014.02.010.

[79] Kiribuchi K, Jikumaru Y, Kaku H, Minami E, Hasegawa M, Kodama O, et al. Involvement of the basic helix-loop-helix transcription factor RERJ1 in wounding and drought stress responses in rice plants. Biosci Biotechnol Biochem 2005. https://doi.org/10.1271/bbb.69.1042.

[80] Yong Y, Zhang Y, Lyu Y. A stress-responsive NAC transcription factor from tiger lily (LLNAC2) interacts with lldreb1 and LLZHFD4 and enhances various abiotic stress tolerance in arabidopsis. Int J Mol Sci 2019. https://doi.org/10.3390/ijms20133225.

[81] Liu C, Wang B, Li Z, Peng Z, Zhang J. TsNAC1 is a key transcription factor in abiotic stress resistance and growth. Plant Physiol 2018.

https://doi.org/10.1104/pp.17.01089.

[82] Zhang Y, Li D, Wang Y, Zhou R, Wang L, Zhang Y, et al. Genome-wide identification and comprehensive analysis of the NAC transcription factor family in Sesamum indicum. PLoS One 2018. https://doi.org/10.1371/journal.pone.0199262.

[83] Tran LSP, Nakashima K, Sakuma Y, Simpson SD, Fujita Y, Maruyama K, et al. Isolation and functional analysis of arabidopsis stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. Plant Cell 2004. https://doi.org/10.1105/tpc.104.022699.

[84] Jeong JS, Kim YS, Baek KH, Jung H, Ha SH, Choi Y Do, et al. Root-specific expression of OsNAC10 improves drought tolerance and grain yield in rice under field drought conditions. Plant Physiol 2010. https://doi.org/10.1104/pp.110.154773.

[85] Song SY, Chen Y, Chen J, Dai XY, Zhang WH. Physiological mechanisms underlying OsNAC5-dependent tolerance of rice plants to abiotic stress. Planta 2011. https://doi.org/10.1007/s00425-011-1403-2.

[86] Pandey AS, Sharma E, Jain N, Singh B, Burman N, Khurana JP. A rice bZIP transcription factor, OsbZIP16, regulates abiotic stress tolerance when over-expressed in Arabidopsis. J Plant Biochem Biotechnol 2018. https://doi.org/10.1007/s13562-018-0448-8.

[87] Jiang AL, Xu ZS, Zhao GY, Cui XY, Chen M, Li LC, et al. Genome-Wide Analysis of the C3H Zinc Finger Transcription Factor Family and Drought Responses of Members in Aegilops tauschii. Plant Mol Biol Report 2014. https://doi.org/10.1007/s11105-014-0719-z.

[88] Ayyappan V, Saha MC, Thimmapuram J. Comparative transcriptome profiling of upland (VS16) and lowland (AP13) ecotypes of switchgrass. Plant Cell Rep 2017;36:129–50. https://doi.org/10.1007/s00299-016-2065-0.

[89] Kim SJ, Kim MR, Bedgar DL, Moinuddin SGA, Cardenas CL, Davin LB, et al.

Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in Arabidopsis. Proc Natl Acad Sci U S A 2004. https://doi.org/10.1073/pnas.0307987100.

[90] Eudes A, Pollet B, Sibout R, Do CT, Séguin A, Lapierre C, et al. Evidence for a role of AtCAD 1 in lignification of elongating stems of Arabidopsis thaliana. Planta 2006;225:23–39. https://doi.org/10.1007/s00425-006-0326-9.

[91] Hisano H, Nandakumar R, Wang ZY. Genetic modification of lignin biosynthesis for improved biofuel production. Vitr Cell Dev Biol - Plant 2009. https://doi.org/10.1007/s11627-009-9219-5.

[92] Fu C, Mielenz JR, Xiao X, Ge Y, Hamilton CY, Rodriguez M, et al. Genetic manipulation of lignin reduces recalcitrance and improves ethanol production from switchgrass. Proc Natl Acad Sci U S A 2011. https://doi.org/10.1073/pnas.1100310108.

[93] Shen H, He X, Poovaiah CR, Wuddineh WA, Ma J, Mann DGJ, et al. Functional characterization of the switchgrass (Panicum virgatum) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. New Phytol 2012;193:121–36. https://doi.org/10.1111/j.1469-8137.2011.03922.x.

[94] Dubos C, Le Gourrierec J, Baudry A, Huep G, Lanet E, Debeaujon I, et al. MYBL2 is a new regulator of flavonoid biosynthesis in Arabidopsis thaliana. Plant J 2008. https://doi.org/10.1111/j.1365-313X.2008.03564.x.

[95] Jackson D, Culianez-Macia F, Prescott AG, Roberts K, Martin C. Expression patterns of myb genes from Antirrhinum flowers. Plant Cell 1991. https://doi.org/10.1105/tpc.3.2.115.

[96] Shen H, Poovaiah CR, Ziebell A, Tschaplinski TJ, Pattathil S, Gjersing E, et al. Enhanced characteristics of genetically modified switchgrass (Panicum virgatum L.) for high biofuel production. Biotechnol Biofuels 2013. https://doi.org/10.1186/1754-6834-6-71.

[97] Li Y jie, Wang B, Dong R rui, Hou B kai. AtUGT76C2, an Arabidopsis

cytokinin glycosyltransferase is involved in drought stress adaptation. Plant Sci 2015. https://doi.org/10.1016/j.plantsci.2015.04.002.

[98] Sallam A, Alqudah AM, Dawood MFA, Baenziger PS, Börner A. Drought stress tolerance in wheat and barley: Advances in physiology, breeding and genetics research. Int J Mol Sci 2019. https://doi.org/10.3390/ijms20133137.

[99] TAKENAKA Y, NAKANO S, TAMOI M, SAKUDA S, FUKAMIZO T. Chitinase Gene Expression in Response to Environmental Stresses in *Arabidopsis thaliana* : Chitinase Inhibitor Allosamidin Enhances Stress Tolerance. Biosci Biotechnol Biochem 2009;73:1066–71. https://doi.org/10.1271/bbb.80837.

[100] Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 2009;25:1105–11. https://doi.org/10.1093/bioinformatics/btp120.

[101] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012. https://doi.org/10.1038/nprot.2012.016.

[102] Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 2015. https://doi.org/10.1093/bioinformatics/btu638.

[103] Filloux C, Cédric M, Romain P, Lionel F, Christophe K, Dominique R, et al. An integrative method to normalize RNA-Seq data. BMC Bioinformatics 2014;15:1–11. https://doi.org/10.1186/1471-2105-15-188.

[104] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:1–21. https://doi.org/10.1186/s13059-014-0550-8.

[105] Du Z, Zhou X, Ling Y, Zhang Z, Su Z. agriGO: A GO analysis toolkit for the agricultural community. Nucleic Acids Res 2010;38:64–70. https://doi.org/10.1093/nar/gkq310.

[106] Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing

biological themes among gene clusters. Omi A J Integr Biol 2012. https://doi.org/10.1089/omi.2011.0118.

[107] Lohse M, Nagel A, Herter T, May P, Schroda M, Zrenner R, et al. Mercator: A fast and simple web server for genome scale functional annotation of plant sequence data. Plant, Cell Environ 2014;37:1250–8. https://doi.org/10.1111/pce.12231.

[108] Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, et al. PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 2017;45:D1040–5. https://doi.org/10.1093/nar/gkw982.

Chapter 3

BUILDING A TEXT MINING PIPELINE TO RETRIEVE LITERATURE TO STUDY STRESS RESPONSE IN ARABIDOPSIS

Hayford R., Arighi C., and Wu C. [version 1; not peer reviewed]. *F1000Research* 2020, **9**(ISCB Comm J):921 (poster) (https://doi.org/10.7490/f1000research.1118198.1)

This chapter is a preliminary work to establish a pipeline to study stress response in plants (discussed in Chapter 4). The Chapter describes an evaluation of available resources for the pipeline and review of related works on plant stress. The goal is to connect resources to study plant stress response. *Arabidopsis thaliana* has been the most widely studied model for all biological processes by the plant science community. Large scale experiments have been conducted using this model to analyze responses and adaptations of plants to environmental stresses.

3.1 Abstract

Environmental stress factors, such a drought and heat, severely affect crop yield. Plants produce a wide variety of responses to endure environmental stress, such as change in rate of photosynthesis. Given the importance of this topic in agriculture, the number of studies is increasing, and so are the publications. Automatically mining information on plant genes and stress could greatly assist biologists conducting research in plant tolerance to stress. Thus, we have established a pipeline to integrate text mining methods to efficiently retrieve information on stress genes and their relation to function and processes in Arabidopsis. For this pipeline, we used

Textpresso, EuroPMPC (ePMC) annotations for gene ontology and GenRIF, and annotations provided by PubTator and pGenN. Upon initial review of 428 abstracts related to Arabidopsis genes and stress, we were able to identify 215 genes and related GO term biological processes from 197 of these. This exercise revealed pain points in the pipeline that need to be improved.

3.2 Background

Text mining is the process of extracting meaningful information from a text which is usually through automated processing of the text [1]. To analyze data correctly and to find the hidden patterns in our data, we need to extract useful information from the data we have. Text mining helps to answer specific research questions, it filters a large amount of research and extracts the relevant information. It can identify and match patterns and trends across millions of articles. This can help to determine additional research that is needed to answer our research question. Text mining helps to draw inferences by combining information from multiple sources [2]. Bioinformatics databases use text mining tools to accurately identify new entries-an example is MirTarBase [3] for validating experimentally microRNA interactions. The applications of text mining are not limited to the manual curation of biological data [4] data integration, gene network interaction and for annotation process [5]. The use of text mining tools for automatic extraction of structural information has heightened due to the rapid growth of biomedical literature and life sciences literature. With the rapid growth of biomedical literature, text-mining tools have attracted more research interests as they can extract structural information from text automatically.

A major challenge in plant research is how plants adapt to climate change. They are frequently exposed to various environmental stresses affecting their growth and development. These environmental stresses have been classified into biotic (e.g. biotrophic and necrotrophic fungi, bacteria, phytoplasmas, oomycetes and nematodes, and non-cellular pathogens i.e. viruses and viroid, pest (phytophagous insects, acari, or nematodes), insects) stress and abiotic stress such as drought, extreme temperatures (cold and heat), extreme light levels, flooding, nutrient deficiency, salinity, chemical factors (heavy metals and pH) and ozone stress [6]. Both biotic and abiotic stresses are major causes of losses in crop yield. Plants, however, have developed complex mechanisms including transcriptomic changes combined with epigenetic regulation to adapt to the stresses, usually depending on the modulation of transcriptional activity of stress-related genes. Various studies have explored and identified stress-responsive genes regulated by stress factor [7,8] which keep increasing. This can lead to relevant data hidden in the text which needs to be extracted automatically. Using text mining methods helps to complement the already known resources with additional information in the area of study. Similarly integrating information available in the specific domain creates a bottleneck and using text mining will enable us to link the necessary information in the literature to specific databases and Ontologies [2] and Textpresso Central [9]

Landeghem et al. (2013) reported on the evaluation of extraction of complex events from literature in plants especially using articles on Arabidopsis [1]. There have been some efforts on developing text mining systems for retrieval of literature relevant to plant research domains, such as PLAN2L system .(pGenN) specifically for plants has also been developed [10]. Text mining techniques have been used to extract

information on the health benefits of medicinal plants and for the extraction of information on disease curing phytochemical properties of medicinal plants [11]. A recent report on the corpus of plant-disease relations in the biomedical domain has been reported by Kim et al [12]. Our goal is to link stress-responsive genes from Arabidopsis to their function (biological processes). To the best of our knowledge, our proposed study will be the first to specifically report a pipeline to extract knowledge on stress genes and their relationship with biological processes in Arabidopsis. We propose to use a controlled vocabulary such as Gene Ontology terms to extract information linking plant genes and stress. The results from the full-scale text mining will be an important resource for researchers especially curators and plant biologists. The pipeline generated can be integrated with PgenN, plant gene normalization tool and other text mining tools.

3.3 Materials and methods

3.3.1 Text mining tools/resources used in the study

We evaluated existing tools and related work for this study. In this section, we first identify the text mining tools that can be used to extract the relevant literature and to set up a pipeline that could support our work in plant stress response. The text mining tools identified for this study were Textpresso [9], PgenN [10], EuroPMC annotations API [14] and PubTator [15].

3.3.1.1 Textpresso Central

Textpresso Central is a text mining system that allows for literature search using keywords and by category (group of terms) using controlled vocabulary. There are three ways to conduct a search on Textpresso: i) Just like regular search engines,

combination of words or phrases; ii) by selecting one or more ontology-based categories from cascading menus; or iii) by combining keyword(s) and categories. Most importantly, Textpresso has tailored corpora, such as Arabidopsis corpus, which was utilized for our search

3.3.1.2 PgenN

PgenN (gene normalization tool): which automatically identifies plant gene names in abstracts and link them to a database entry. PgenN is a plant gene dictionary-based tool which was created based on plant proteins in the UniProt database [10]text mining tool was used to retrieve plant genes which was compared to the output from pgenN [15].

3.3.1.3 PubTator

PubTator: Also, we used PubTator to retrieve bioconcept annotations and this text mining tool was used to retrieve plant genes which was compared to the output from pgenN [15].

3.3.1.4 EuroPMC

EuroPMC: ePMC offer annotations based on text mining approaches to retrieve comprehensive annotations from publications [14] We used the services to retrieve functional annotation and terms that matches to GO terms.

3.3.2 Retrieval of publication

We called the API of Textpresso from the terminal [34] using the script below from the *A. thaliana* corpus.

(curl -k -d "{\"token\":\"fcskR8ayJjJSGGsWdLQ2\", \"query\": {\"keywords\": \"abiotic AND stimulus\", \"type\": \"document\", \"corpora\": [\"PMCOA A.

thaliana\"]}}"
https://textpressocentral.org:18080/v1/textpresso/api/search_documents.), date accessed-September 2020). Different keywords like " "abiotic AND stimulus" and "biotic AND stimulus" were used to retrieve publications from the Arabidopsis corpus on Textpresso. This was done using the command line. The output response which is a JSON format include information such as an accession (PMID) for the document, the title, the authors and a score of the document which defines the extent to which the document matches with the query. The PMIDs retrieved from Textpresso were used as an input to obtain gene mentions from the abstracts using PgenN and PubTator. The aim was to compare and incorporate the results from pgenN and PubTator. HES-SO/SIB Text Mining tool for Elixir through EuroPMC provided the annotation for the functional sentences (which are mappings of GenRIF sentences to their corresponding text). Similarly, the terms that matches to GO terms associated with the plant genes identified were annotated for each of the selected publications using ePMC. The output of the functional sentences was evaluated to determine if there is a link or relation between the GO term and the gene. A procedure for retrieving publications, collecting plant stress-related genes, and annotation of functional sentences and GO terms has been depicted in **Figure 3.1.**

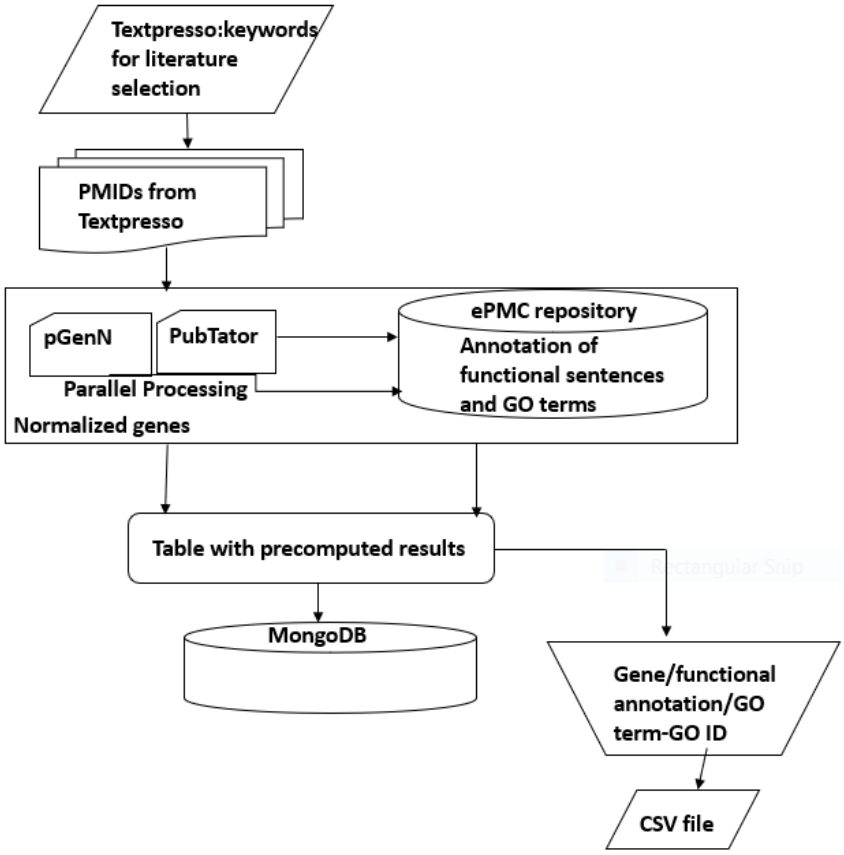


Figure 3.1: Workflow to retrieve genes and link them to their function in Arabidopsis

3.4 Results and discussion

3.4.1 Basic statistics from data collected

In this section, the statistics of the data collected was presented. Using different keywords in both the "document and sentence" scope from the Arabidopsis corpus of Textpresso, we retrieved 815 unique PMIDs of publications related to biotic and abiotic stress on September 20, 2019. Although, there were about 5000 stress-related publications that were recorded from PubMed for the 2000-2019 we only retrieved 815 PMIDs from Textpresso. This is because we only accessed the stress-

related publications from the Arabidopsis corpus which excluded other plants. Out of the 815 publications, 428 and 758 publications with stress-related genes were retrieved from pgenN and PubTator, respectively. The data downloaded from PubTator captured any species with 396 PMIDs indicated for only Arabidopsis. A total of 727 stress-related genes were mentioned in pgenN (sheet #5 of supplemental data). However, from PubTator using the same PMIDs, 1333 unique plant stress-related genes were identified (sheet #7 of the supplemental data). We collected annotations of 110 genes out of the 727 genes from pgenN with functional sentences and GO terms. The PMIDs with no functional sentences were completed with GO terms (sheet #11 of supplemental data). The 110 functional annotations were evaluated to determine if there is a relation between the gene mentioned and the GO term (those with the biological process domain were selected). We identified 51 annotations with the suggested relation between the gene mention and GO term and functional sentence . Also, 21 annotations with gene and related GO terms within the same functional sentence as shown in Fig. 3.3. This group of classification (annotation) serve as high confidence group. A summary of the data collected has been summarized in **Table 3.1 and Figure 3.2**. A pre-computed table includes PMIDs, gene mentioned from pgenN, PubTator, UniProt accession, sentence annotation, functional annotation, provider for functional annotation, GO term annotation and GO term annotation ID. A section indicated in Figure 3.4 and Figure 3.5.

Table 3.1: Summary of the statistics of data collected on plant stress genes and GO terms

Abstracts	**No. of abstracts with genes detected**		**No. Gene mentions**		**No. abstracts with annotation of functional sentences**	**No. abstracts with gene-GO term relation(out of the no. of PgenN abstracts)**
	PgenN (plant)	PubTator (any species)	PgenN	PubTator	EuroPMC	EuroPMC
815	428	758	725	945	110	197

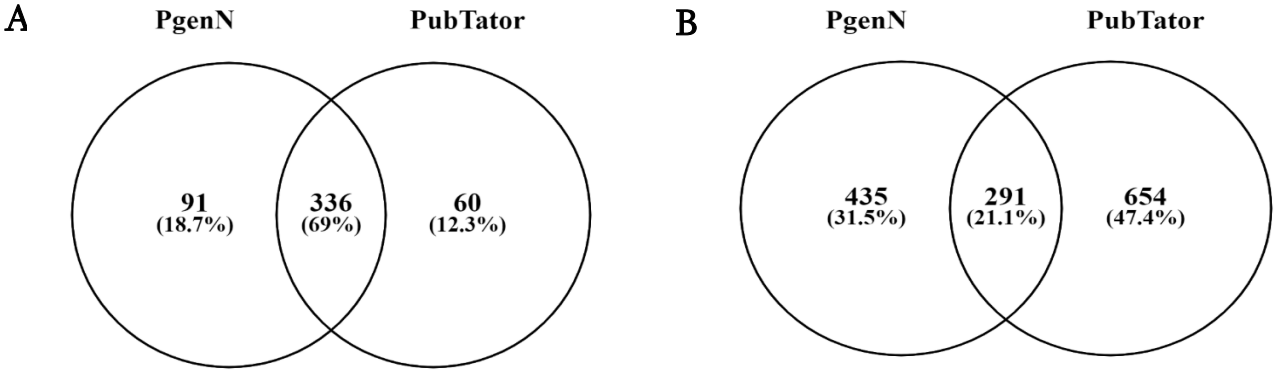


Figure 3.2: Unique abstracts and gene mention between PgenN and PubTator. A) number of abstracts. B) number of gene mention. We used PgenN and PubTator complementing each other to retrieve gene mentions.

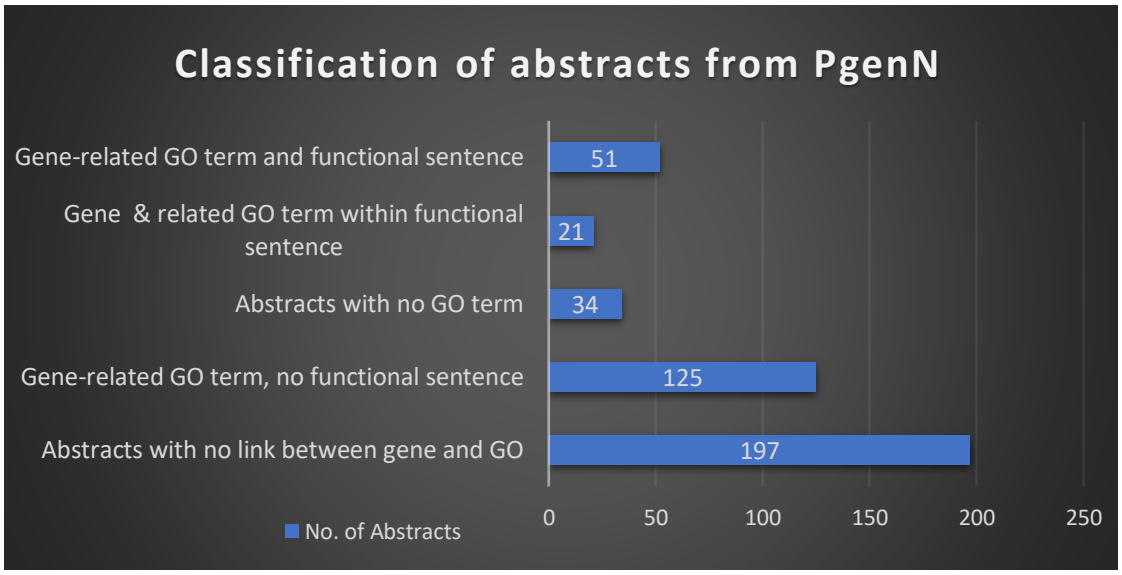


Figure 3.3: Classification of abstracts based on gene biological process (GO term) relationship. The abstracts pulled from PgenN were used for the classification. Manual inspection of the data was conducted to classify the abstracts. The precomputed table is available via this link; (https://docs.google.com/spreadsheets/d/1F1joXBbIWPYEhKyiVNIYPIftsFKCUAWn/edit?usp=sharing&ouid=105671503806176794393&rtpof=true&sd=true)

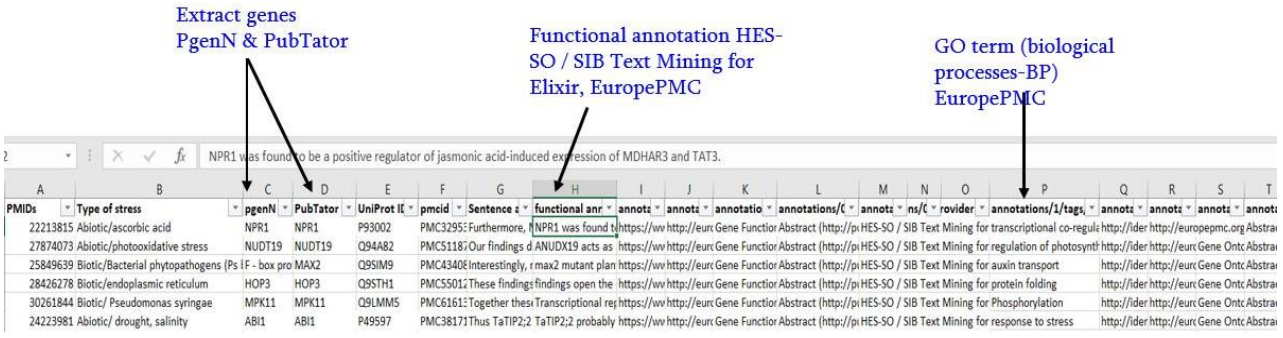
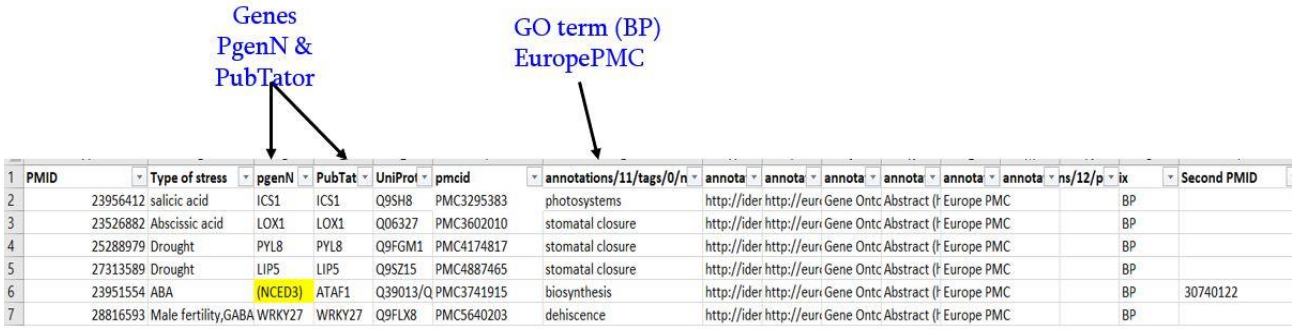


Figure 3.4: Shows a section of the manual curated data with gene mention, functional annotation and related GO terms. High confidence of gene mention and GO term (BP) within the functional annotated sentence.



	PMID	Type of stress	pgenN	PubTat	UniProt	pmcid	annotations/11/tags/0/n	annota	annota	annota	annota	annota	annota	ns/12/p	ix	Second PMID
2	23956412	salicic acid	ICS1	ICS1	Q9SH8	PMC3295383	photosystems	http://ider	http://eur	Gene Ontc	Abstract (h	Europe PMC			BP	
3	23526882	Abscissic acid	LOX1	LOX1	Q06327	PMC3602010	stomatal closure	http://ider	http://eur	Gene Ontc	Abstract (h	Europe PMC			BP	
4	25288979	Drought	PYL8	PYL8	Q9FGM1	PMC4174817	stomatal closure	http://ider	http://eur	Gene Ontc	Abstract (h	Europe PMC			BP	
5	27313589	Drought	LIP5	LIP5	Q9SZ15	PMC4887465	stomatal closure	http://ider	http://eur	Gene Ontc	Abstract (h	Europe PMC			BP	
6	23951554	ABA	(NCED3)	ATAF1	Q39013/Q	PMC3741915	biosynthesis	http://ider	http://eur	Gene Ontc	Abstract (h	Europe PMC			BP	30740122
7	28816593	Male fertility,GABA	WRKY27	WRKY27	Q9FLX8	PMC5640203	dehiscence	http://ider	http://eur	Gene Ontc	Abstract (h	Europe PMC			BP	

Figure 3.5: Shows a section of the manual curated data with gene mention and related GO term.

3.5 Challenges faced in the study

At the time of the study, one of the limitations was how PubTator annotated stresses in plants in the scientific literature. Since PubTator has high recall there was also the issue of redundancy in the data collected. Also, the task of associating an entity to an experimental evidence tend to be a challenge as elaborated or mentioned by Hirschman et al. [16]. One of the pitfalls was the fact that GeneRIF (ePMC) annotations were not extensive for Arabidopsis. Hence not all the functional annotations were retrieved for most of the normalized genes.

3.6 Next steps

For the next steps, the most important one was to extend the pipeline to retrieve information on all plants and make the pipeline systematic, run automatically and store the output in MongoDB. In addition, we explored other text mining methods that will enable us to enrich the bibliome content related to stress response in Arabidopsis and plants in general. The data collected was consolidated to the data obtained in chapter 4.

REFERENCES

[1] Van Landeghem S, De odt S, Drebert ZJ, Inzé D, Van De Peer Y. The potential of text mining in data integration and network biology for plant research: A case study on Arabidopsis. Plant Cell 2013;25:794–807. https://doi.org/10.1105/tpc.112.108753.

[2] Krallinger M, Valencia A, Hirschman L. Linking genes to literature: Text mining, information extraction, and retrieval applications for biology. Genome Biol 2008. https://doi.org/10.1186/gb-2008-9-s2-s8.

[3] Chou CH, Chang NW, Shrestha S, Hsu S Da, Lin YL, Lee WH, et al. miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res 2016. https://doi.org/10.1093/nar/gkv1258.

[4] Arighi CN, Lu Z, Krallinger M, Cohen KB, Wilbur WJ, Valencia A, et al. Overview of the BioCreative III Workshop. BMC Bioinformatics 2011;12:S1. https://doi.org/10.1186/1471-2105-12-S8-S1.

[5] Van Landeghem S, De Bodt S, Drebert ZJ, Inzé D, Van De Peer Y. The potential of text mining in data integration and network biology for plant research: A case study on Arabidopsis. Plant Cell 2013. https://doi.org/10.1105/tpc.112.108753.

[6] Nejat N, Mantri N. Plant immune system: Crosstalk between responses to biotic and abiotic stresses the missing link in understanding plant defence. Curr Issues Mol Biol 2017. https://doi.org/10.21775/cimb.023.001.

[7] Cohen SP, Leach JE. Abiotic and biotic stresses induce a core transcriptome response in rice. Sci Rep 2019;9:1–11. https://doi.org/10.1038/s41598-019-42731-8.

[8] Li X, Li M, Zhou B, Yang Y, Wei Q, Zhang J. Transcriptome analysis provides insights into the stress response crosstalk in apple (Malus × domestica) subjected to drought, cold and high salinity. Sci Rep 2019. https://doi.org/10.1038/s41598-019-45266-0.

[9] Müller HM, Van Auken KM, Li Y, Sternberg PW. Textpresso Central: A customizable platform for searching, text mining, viewing, and curating biomedical literature. BMC Bioinformatics 2018. https://doi.org/10.1186/s12859-018-2103-8.

[10] Ding R, Arighi CN, Lee JY, Wu CH, Vijay-Shanker K. pGenN, a gene normalization tool for plant genes and proteins in scientific literature. PLoS One 2015;10:1–23. https://doi.org/10.1371/journal.pone.0135305.

[11] Behera NK, Mahalakshmi GS. A cloud based knowledge discovery framework, for medicinal plants from PubMed literature. Informatics Med Unlocked 2019;16:100105. https://doi.org/10.1016/j.imu.2018.04.006.

[12] Kim B, Choi W, Lee H. A corpus of plant–disease relations in the biomedical domain. PLoS One 2019. https://doi.org/10.1371/journal.pone.0221582.

[13] Ren J, Li G, Ross K, Arighi C, McGarvey P, Rao S, et al. iTextMine: integrated text-mining system for large-scale knowledge extraction from the literature. Database (Oxford) 2018;2018:1–10. https://doi.org/10.1093/database/bay128.

[14] Levchenko M, Gou Y, Graef F, Hamelers A, Huang Z, Ide-Smith M, et al. Europe PMC in 2017. Nucleic Acids Res 2018. https://doi.org/10.1093/nar/gkx1005.

[15] Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res., vol. 47, Oxford University Press; 2019, p. W587–93. https://doi.org/10.1093/nar/gkz389.

[16] Hirschman L, Burns GAPC, Krallinger M, Arighi C, Cohen KB, Valencia A, et al. Text mining for the biocuration workflow. Database 2012;2012:1–10. https://doi.org/10.1093/database/bas020.

Chapter 4

A PIPELINE TO AUTOMATICALY RETRIEVE INFORMATION ON PLANT STRESS TO SUPPORT ANNOTATION OF SWITCHGRASS (PANICUM VIRGATUM L.)

(Hayford R.K, Arighi, C.N, Kalavacharla V.& Wu, C.H., (In Review 2022))

4.1 Abstract

Environmental stresses such as drought, extreme temperatures, salinity, and pathogens threaten global crop productivity. Therefore, developing cultivars with improved tolerance to such stresses has emerged as the most sustainable solution and an area of active research. The literature contains a wealth of information about plant biology that can be harnessed to improve gene annotation, experimental design, and hypothesis generation for non-model organism species. This study used computational and experimental approaches to understand better plant stress response mechanisms in switchgrass, a critical bioenergy crop. Thus, we established a pipeline integrating data from databases and text mining methods to efficiently retrieve publications relevant to plant stress genes and link them to their function from the Scientific literature. This pipeline uses literature and annotations from several sources, including Medline, Textpresso, pGenN, UniProt, and ePMC, and co-occurrence of a stress gene and an annotation in the same abstract. The data is stored in a MongoDB database currently containing 2,766 abstracts, 3,716 unique plant stress-responsive genes, 861 GO terms, and 1,007 GenRIF sentences. We used the MongoDB collection to check against a list of stress-responsive genes from switchgrass RNA-Seq data to see if the stress gene is

over or under-expressed. Other GO biological processes of the switchgrass stress genes were identified based on homology-based inference from the literature collected. One interesting candidate gene that encodes Phenylalanine ammonia-lyase 1 (PAL1) was selected for experimental validation. PavirPAL1 showed PAL activity at a temperature of 30 ^{o}C and pH 8.5. Our results indicate that the database developed in this study could support gene-annotation enrichment tools to provide the function of plant stress genes from high-throughput gene expression data. Automatically mining information on plant genes and stress could greatly assist biologists in researching plant tolerance to stress.

4.2 Introduction

Plants are constantly exposed to various environmental stress factors (biotic and abiotic), which severely affect crop productivity worldwide, leading to their loss of up to 70% [1], [2]. Under natural conditions, multiple stresses occur: the effects of the biotic and abiotic stresses may occur singularly or in combination, inducing enumerable damages at different developmental stages of plants [3], [4]. As sessile organisms, plants cannot escape the environmental stress conditions; therefore, they adapt to these conditions by developing complex mechanisms, including signaling pathways transcriptome changes, to tolerate these stresses [5], [6], [7]. Plants produce a wide variety of responses to enduring environmental stress, such as a change in the rate of photosynthesis. Developments in the OMICs technology over the past decade have provided the platform to conduct complex studies to understand the molecular mechanisms underlying stress responses in plants. For example, transcriptome approaches have revealed several differentially regulated genes during normal and under stress conditions. In addition to the differentially expressed genes, the high-

throughput experiments have shown information such as gene ontology terms (standard to describe gene function) and pathway analysis, reported in the scientific literature.

There are a number of resources developed with data on stress response genes in plants. Borkotoky et al. (2013)[8] have reported on a collection of data on stress-responsive genes in a database specific for Arabidopsis. The STIFDB2 database [3]is an updated version of the plant stress-responsive transcription factor database covering three species: *Arabidopsis thaliana* and *Oryza sativa* subsp. Japonica and *Oryza sativa* subsp. Indica.STIFDB2 integrates data mining of genomic data, biocuration, and prediction to collect TFs and stress-responsive genes. The web interface is active but was last updated in 2012. Further, a plant stress gene database (PSGD) containing information on stress genes and their ortholog and paralog in plants has been reported [9]. The PSGD was created through literature and database search using keywords. PSGD integrates information from databases such as NCBI and EMBL. The web interface for PSGD is active but was last updated in 2011. Similarly, a database for plant proteome response to stress (PlantPRes) that contain manually curated articles on stress proteins in plants has been reported [10]. The PlantPRes database interface is active and recently updated. In the same line, a database of rice transcription factors under stress conditions was created using data from the Plant TF database[11]. The web interface for the RiceSRTFDB is not accessible. A database with a primary focus on drought stress, the DroughtDB, has been reported to assist researchers working on drought stress in plants [12]. The droughtDB was developed based on a manual compilation of drought genes that are molecularly characterized, it was last updated in 2014.

Given the importance of this topic in agriculture, the number of studies on plant stress keeps increasing (Fig. 1). These studies, which can understand what is happening at a global, whole genome-scale in response to external stimuli such as stress, can also produce vast amounts of data that can be difficult to analyze to derive meaningful conclusions [13]. Using text mining approaches will help extract structural information from the text. Using text mining systems increases knowledge extraction to complement already known resources with additional information to our study area and assist hypothesis generations.

Elucidating the basic biological knowledge behind the stress response mechanisms in plants is vital to devise strategies to improve plant tolerance against stresses [3]. Although databases on plant stress exist, some of these databases are defunct, while others use existing data, and are not literature-based. Here, we describe a method by integrating experimental studies, text mining methods, and computational methods to determine the function of stress-responsive genes. To find relevant data published on plant stress, we developed a pipeline (Fig. 2) that integrates a text mining results from PgenN, a tool that identifies plant genes [14] and literature database [15], namely Europe PMC, that highlights terms that match GO terms and gene to function (GenRIF) sentences. We use the co-reference of terms in the text as a simple method to show a relationship between them. Identifying genes and their co-occurrences with relevant keywords from the literature have been widely used [16], [17], [18]. In our case, we focus on the co-mention of a GO term, functional sentences, and stress gene in the same sentence and/ or paragraph. We want to increase the chance that the publication describes experiments about the gene's possible involvement in such a process if the GO term is for biological processes. We hypothesize that the literature can supplement homology-based inference of the processes of differentially expressed genes from an under annotated species, such as switchgrass, by providing data with

specific context on the experimental stress condition. A database on plant stress genes with related GO terms or functional annotation has been developed. The information collected is used to predict the function of stress-responsive genes identified from RNA-Seq data on switchgrass (*Panicum virgatum* L) imposed with a single drought and combinations of drought and heat stress [19].

Switchgrass is warm, C4 perennial grass, and a critical bioenergy crop. It grows on marginal lands and could help reduce the global energy shortage [20],[21]. Switchgrass is considered a forage crop for livestock, has a high biomass yield, and produces biofuel [22]. The US Department of Energy has named switchgrass a herbaceous biofuel feedstock model [23]. Despite the economic, agricultural, and environmental importance of switchgrass, drought and heat stresses have been limiting factors for switchgrass biomass and biofuel production [22], [24]. Recent studies have been conducted on the assembly and annotation of switchgrass [23] Reports on adequate knowledge of genes to support crop production and adaptation have primarily focused on a small number of well-studied model plants [23]. The function of switchgrass genes will be predicted using information collected from the scientific literature. Using information from the scientific literature helps to identify annotation (e.g., processes) of the genes from orthologs that may not be captured in the existing databases. In addition, this information usually has experimental evidence to support the annotation of the genes. To infer the role of the genes predicted in switchgrass, as a case study, we wanted to understand better the function of a Putative Phenylalanine Ammonia-Lyase 1 (PAL1) protein from switchgrass designed as *PavirPAL1*. Phenylalanine Ammonia-Lyase (PAL, EC 4.3.1.24) is the first enzyme in the phenylpropanoid pathway that catalyzes the deamination of $_{L}$-phenylalanine to trans-

cinnamic acid. It plays a crucial role in plant development and defense [25]. Genes involved in the phenylpropanoid pathway have been revealed to play a significant role in feedstock quality [20]. PAL is one of the most important secondary metabolic pathways in plants. As a result of its role in the phenylpropanoid pathway, it has been extensively studied in many plants, including *Arabidopsis thaliana* [26], *Oryza sativa* [27], *Zea mays* [28], *Solenostemon scutellarioides* [29], *loblolly pine* [30], *salvia miltirrhiza* [25], *lycoris radiata* [31], *Bambusa oldhamii* [32], *Juglans regia* [33], and *Melissa officinalis* [34].

Using information from the MongoDB collection, we identified additional annotations, including cold acclimation (GO:0009631), secondary metabolism (GO:0019748), anthocyanin synthesis (GO:0009718), and biosynthesis (GO:0009058), which are linked to PAL1 from Arabidopsis and *Solenostemon scutellarioides* (coleus). These annotations were not reported in UniProt, TAIR, or in the Gene Ontology Resource at the time of the data collection. In addition, we identified from our literature collection that PAL1 was induced by cold and light stress from Arabidopsis and coleus, respectively. We propose inferring these functions to their ortholog PAL1 gene (transcript name; Pavir.1KG386300.v4.1 referred here as *PavirPAL1*), responsive to switchgrass combined drought and heat stress. To confirm or validate this putative PAL gene, we isolated PAL1 from switchgrass. The recombinant PavirPAL1 was characterized by determining the protein activity by catalyzing $_{L}$-phenylalanine to trans-cinnamic acid. Our study provides a comprehensive approach to identify stress-responsive genes and validate their function. To our best knowledge, the analysis here is the first detailed investigation of

the PAL enzyme of the phenylpropanoid biosynthesis in *Panicum virgatum*. This study will potentially help develop switchgrass cultivars with better tolerance to stress.

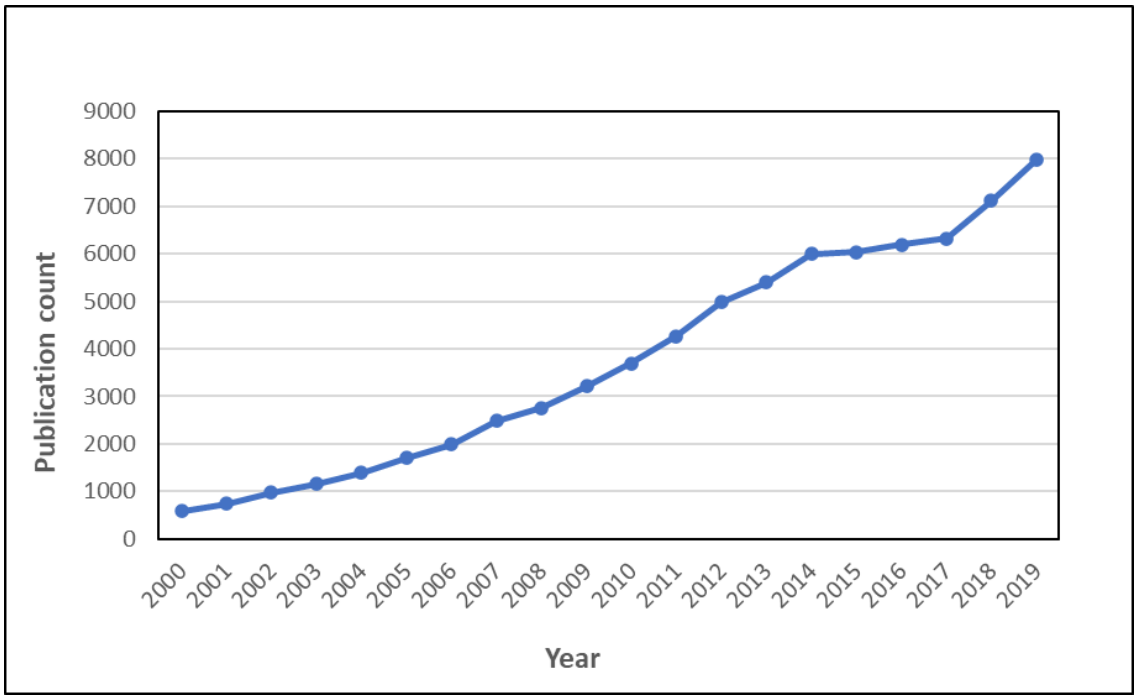


Figure 4.1: Number of scientific publications on stress study in plants from PUBMED database from 2000 to 2019. PudMed database was queried using this script (pubmed - (("YYYY/MM/DD "[Date - Publication] : "YYYY/MM/DD"[Date - Publication])) AND stress) AND "Plants"[MeSH]). The dates for the beginning and end of each year were inserted to retrieve the literature for each year (data was retrieved from PudMed on April 2021)

4.3 Materials and methods

4.3.1 Resources used in the study

This section first identified the resources that can extract the relevant literature and set up a pipeline that could work to retrieve plant stress genes and annotations.

The resources identified for this study are Textpresso [35], PgenN [14], Europe PMC [15,36], and UniProt [37] are other resources included in the pipeline. Textpresso Central is a text-mining tool that allows literature search using keywords and by category (group of terms). Textpresso has corpora such as Arabidopsis corpus, which was utilized for our search. PgenN (gene normalization tool) automatically connects plant gene names to a database. PgenN identifies normalized plant genes and links them to the UniProt database. The UniProt Knowledgebase (UniProtKB) is a publicly available database offering sequence and functional annotation for proteins across all taxonomic groups. We used UniProt for ID mapping and Europe PMC to retrieve functional annotation and GO terms. Europe PMC offers annotations based on text mining approaches to retrieve comprehensive annotations from publications [36]. It is a literature database that highlights terms that match to GO terms.

4.3.2 Description of the pipeline for data collection

Figure 4.2. illustrates an overview of the pipeline to retrieve plant stress genes and link them to their function. The pipeline integrates (PgenN)[14], UniProt knowledgebase [37] (EuroPMC) [15] and information from Textpresso [35]. We want to use the simplest method of co-occurrence of two concepts within the same document, using the abstract. The co-location of a GO term and stress gene provides a higher probability that the publication describes experiments about the gene's possible involvement in such a process if the GO term is for a biological process. The PMIDs of the publications used in the study were retrieved from Textpresso and PudMed. The pipeline undergoes a serial process of parsing the processed abstracts to PgenN to retrieve normalized plant genes, then to UniProt for Id mapping. The abstracts with

normalized plant genes are parsed to Europe PMC to retrieve GO terms and functional annotation sentences. Abstracts with normalized genes and annotations are stored in MongoDB. To enrich our data, we used GO term biological processes related to plant stress to collect genes with the linked publication from the Gene Ontology database.

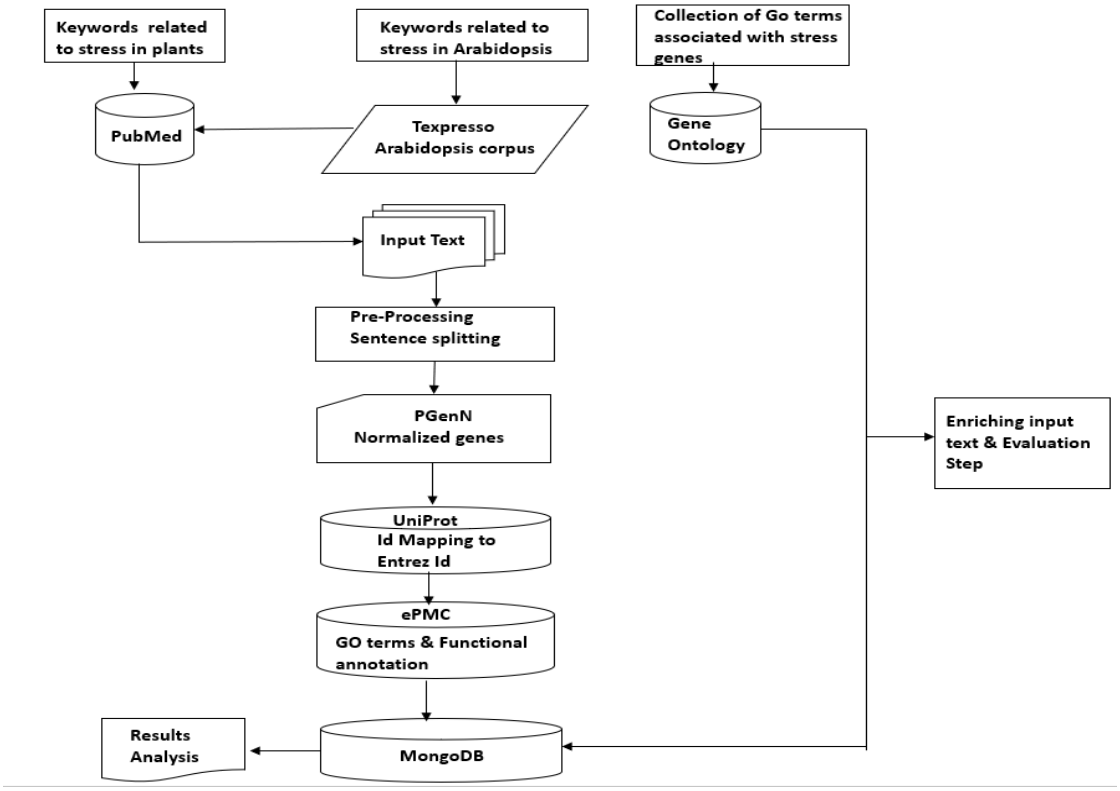


Figure 4.2: Workflow of our plant-stress-gene-annotation relationship extraction

4.3.3 Retrieval of publications on plant stress

We called the API of Textpresso from the terminal [34] using the script below from the *A. thaliana* corpus.

(curl -k -d "{\"token\":\"fcskR8ayJjJSGGsWdLQ2\", \"query\": {\"keywords\": \"abiotic AND stimulus\", \"type\": \"document\", \"corpora\": [\"PMCOA A.

thaliana\"]}}"
https://textpressocentral.org:18080/v1/textpresso/api/search_documents.), date Faccessed-September 2020). Different keywords like " "abiotic AND stimulus" and "biotic AND stimulus" were used to retrieve publications from the Arabidopsis corpus on Textpresso. This was done using the command line. The output response which is a JSON format include information such as an accession (PMID) for the document, the title, the authors and a score of the document which defines the extent to which the document matches with the query. To expand our search of literature to include all plants we used selected keywords with the PubMed API from the E-Utilities documentation [38] using the following scripts;
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Abiotic+AND+Stimulus+AND+Plant[Mesh]+&retstart=1&retmax=120&usehistory=y"
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Abiotic+AND+Stress+(("2000/01/01 "[Date - Publication] : "2018/12/01"[Date - Publication]))+AND+Plant[Mesh]+&retstart=1&retmax=100000&usehistory=y"
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=biotic+AND+Stress+(("2000/01/01 "[Date - Publication] : "2018/12/01"[Date - Publication]))+AND+Plant[Mesh]+&retstart=1&retmax=100000&usehistory=y"
https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=biotic+AND+Stimulus+(("2000/01/01 "[Date - Publication] : "2018/12/01"[Date - Publication]))+AND+Plant[Mesh]+&retstart=1&retmax=100000&usehistory=y"**(The data was retrieved from PubMed on September 28, 2020).** In addition, we collected GO terms (biological process) associated to abiotic and biotic stress [38,39], and used them as input in the Gene Ontology Resource to retrieve publications and stress genes linked to these terms. The abstracts retrieved from PudMed were used as input text in the pipeline.

4.3.4 Text preparation and data generation

The abstracts (as input text in Medline abstract format) for processing were first split into individual sentences using an in-house developed tool in JSON format. The input text is loaded and iterates through the split document. Each PMID entry is parsed through PgenN to retrieve normalized plant genes and UniProt for UniProt accession and Entrez ID mapping [41]. Next, terms that match GO terms and annotation for the functional sentences (which are mappings of GenRIF sentences to their corresponding text) were retrieved using HES-SO/SIB Text Mining tool for Elixir through Europe PMC. If the parsed response has a gene mention and annotation, it saves directly in the MongoDB database in a JSON format. Duplicate genes were removed from the MongoDB using the unique UniProt accession of the plant stress genes.

4.3.5 Evaluation of the pipeline

We used the GO annotation resource as an evaluation corpus for our pipeline and a method to enrich the database. GO annotations (biological process) selected from publications on abiotic and biotic stress response [39,40] were used to query GO by filtering the annotations using “Viridiplantae” as organism and annotations with experimental based evidence. This process ensured adequate coverage of literature on stress genes with annotations in our database. We confirmed some literature with plant normalized genes and annotations retrieved with our pipeline. Using GO resources to enrich the database, we identified additional publications with stress genes and GO terms, this information was added manually to the MongoDB.

4.3.6 Validation of the MongoDB database on plant stress genes and annotation

To validate usability of the MongoDB collection, we evaluated the possibility of using the collection of publications in this study to search for evidence of the role of differentially expressed genes identified from switchgrass transcriptomes when imposed with single drought and combination of drought and heat stress [19]. The switchgrass RNA-Seq data was inspected to identify differentially expressed genes with annotation of orthologs not curated in the Gene Ontology Resource but has annotation linked to them from the publications in our collection. To infer the predicted annotations to the switchgrass stress-responsive genes, we further used computational/bioinformatic analysis of PAL1 orthologs (*Arabidopsis thaliana* and *Solenostemon scutellarioides* (coleus)) to determine its similarity with switchgrass PAL1 *PavirPAL1*.

4.3.7 Sequence and bioinformatic analysis

The coding sequence of *PavirPAL1* was obtained from Phytozome, a plant sequence database where the switchgrass genome is deposited [42]. To determine the percentage identity of *PavirPAL1* with its orthologs, we obtained the protein sequences of the orthologs from the UniProt database (UniProt Consortium, 2017) and BLAST done against the switchgrass version 4 genome [43]. Multiple sequence alignment was made using the ClustalW algorithm of DNAMAN software-version 10 (Lynnon BioSoft Coporation, Quebec, Canada) ([44]. A detailed analysis of *PavirPAL1* was conducted using computational tools and comparing the output with orthologs in Arabidopsis and coleus (publications on *AtPAL1* and *SsPAL1* contain annotations which will be inferred). DeepLoc-1.0 [45] prediction algorithm was used

to predict the subcellular localization using only the sequence information. The physiochemical properties like the molecular weight and isoelectric point (pI) were theoretically determined using the ProtPram tool in ExPASy [46]. The conserved domain, motifs and family of *PavirPAL1* were predicted using the Conserved Domain Database (CDD) [47], Conserved Domain Architecture Retrieval Tool (CDART) [48], ScanProsite [49], motif finder [50], and InterProScan [51]. The transcriptional start site and putative cis-acting elements were predicted using Plant Cis-acting Regulatory DNA Elements (PLACE) [52] and Plant-Care database [53]. To predict the tertiary structure of *PavirPAL1*, the online tool SWISS-MODEL [54] was used to determine the 3-dimensional (3D) structure. A known homolog of *PavirPAL1*from *Petroselium crispsum* (*PcPAL*, Iw27.1.A) [29,55] from the protein data bank (PDB) was used as a template to build the 3D structure of *PavirPAL1*. The secondary structure was predicted using the SOPMA program[56].

4.3.8 Molecular characterization of *PavirPAL1*

4.3.8.1 *Expression of PavirPAL1 in E.coli and purification of recombinant proteins*

Analyzing the in-house RNA-Seq data on switchgrass showed three drought, and heat-responsive genes encoded PAL1. However, the gene with transcript name Pavir.1KG386300.v4.1 (*PavirPAL1*) was selected due to its response to combined drought and heat stress at two different points (DTHT 96/24 h and DTHT 168/96 h). The other two transcripts Pavir.1NG356800.v4.1 and Pavir.7NG355800.v4.1, were differentially expressed at time point 168/96 h. The full-length coding sequence encoding of *PavirPAL1* with transcript ID Pavir.1KG386300.v4.1 was codon-optimized and synthesized by GenScript (Piscataway, NJ, USA). The codon-optimized

putative *PavirPAL1* gene (Pavir.1KG386300.v4.1) was cloned into the pET30a. The pET30a-*PavirPAL1* was constructed by inserting *PavirPAL1* into the *Ndel-HindIII* site of pET30a. A $(His)_6$-tag was fused to the recombinant protein to facilitate protein purification. The pET30a-*PavirPAL1*was transferred into *E.coli* BL21 Star™ (DE3) competent cells for protein expression. A single colony was inoculated into LB medium containing kanamycin; cultures were incubated in 37 ℃ at 200 rpm. Once cell density reached OD=0.6-0.8 at 600 nm, 0.5 mM IPTG was introduced for induction. SDS-PAGE and Western blot were used to monitor the expression (Figure 7). The protein was obtained from the cell lysate supernatant and purified using TALON Metal Affinity Resin+Superdex 200.

4.3.8.2 Analytical methods/enzymatic assay test/ biochemical assay

The enzymatic activity of PAL was assayed from the purified recombinant protein by modifying previous assay report [57,58]. The PAL activity was performed by measuring the formation of trans-cinnamic acid from L-phenylalanine. Phenylalanine Ammonia-Lyase, from *Rhodotorula glutinis* sigma, P1016, 1.2 U/mg (https://www.sigmaaldrich.com/NL/en/product/sigma/p1016?context=product) was used as a standard. The dilution buffer for the reaction contained 10 mM Tris, 200 mM NaCl, 10% glycerol at pH8.5 with a Tris solution at 360 mM Tris-HCl, pH 8.8 and L-Phenylalanine solution, 72 mM L-Phenylalanine. The standard sample was diluted 200 times to obtain a diluted concentration for activity test and three different serial dilutions for the test sample, recombinant Pavir.1KG386300.1 or *PavirPAL1* (Supplemental table S4) . The reaction was incubated at a temperature of 30℃ and wavelength = 270 nm. The absorbance was read every minute for 10 minutes in total. A linear equation for the change in absorbance of samples of different concentrations

within the 10 minutes was calculated. The change in absorbance value for the standard sample was used to calculate the activity of the test sample. The enzyme assay was performed in triplicate.

4.3.9 Validation of RNA-Seq of *PavirPAL1*(Pavir.1KG386300.v.1) using traditional and qRT-PCR

Primers were designed to conduct traditional PCR and quantitative real-time PCR to validate the expression of *PavirPAL1* (Pavir.1KG386300.v4.1) from the switchgrass RNA-Seq data. The primers (Table S2) were designed using the online tool for real-time PCR (TaqMan) primer design by GenScript (Piscataway, NJ USA). Total RNA was isolated from the same plant materials used for the switchgrass RNA-Seq experiments [19] using RNeasy Plant Mini Kit (Qiagen, Inc., Germany) following the manufacturer's protocol. The concentration and purity of the RNA samples were determined using a Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA). The integrity of all RNA samples was evaluated by gel electrophoresis. First-strand cDNA was synthesized using DNase treated RNA using Postscripts II cDNA synthesis kit (New England BioLabs, Ipswich MA). Conventional PCR was performed using 1 µl of 100 ng of cDNA as a template for the traditional PCR reaction under these conditions (94^{o}C for 1 min, 60^{o}C for 30 s, and 72^{o}C for 1 min) for 30 cycles. The PCR product was separated on a 1% agarose gel stained with ethidium bromide.

QRT-PCR was performed using the synthesized cDNA. Twenty-five µLs of the PCR reactions containing1 100ng of 1st-strand cDNA, 12.5 µL of Power SYBR Green Master Mix, and 3 µLs of 10 nM specific primers (forward and reverse) and 9.5 µL of water. We used cons7, primer sequence is given (Table S2)a as a reference

gene or internal control primer to normalize the results in all the samples [59]. The PCR conditions for the qRT-PCR were the following; 95ºC for 10 min, followed by 40 cycles of 95ºC for 15 s and 65ºC for 1 min. The qRT-PCR was performed using an ABI 7500 real-time PCR system and SYBR Green Kit (Applied Biosystems, Grand Island, USA). The relative expression was determined from three biological and two technical replicates using ΔΔCT method [60]. Minitab- software (State College, PA) was used to analyze the normalized CT values from the qRT-PCR analysis.

4.4 Results and discussion

4.4.1 Database generation

We developed a pipeline to automatically mine information on plant genes and stress from the scientific literature. PubMed and Textpresso were used as a resource to collect publications. Using terms related to stress, as mentioned in the methods section, we were able to retrieve specific abstracts for processing. Gene ontology terms related to plant stress were used to query GO resources to retrieve publications and plant stress genes with experimental evidence. PgenN automatically identifies plant gene names in abstracts and link them to a database entry. Our collection contains information retrieved from abstracts since PgenN tool for retrieving normalized gene cannot process full-text information [14]. Europe PMC offers text mining-based annotations to retrieve functional annotations and GO terms from publications [36]. A total of 13,913 Medline was processed using the pipeline developed. The documentation for running the pipeline and is stored in a GitHub repository https://github.com/udel-biotm-lab.

We implemented a database of plant stress genes and their function. The URL for the database hosted locally is mongodb://localhost:27017/database. A description of the dataset of the latest version of the MongoDB collection is indicated in **Table 4.1**. In addition to the statistics of the content of the database, the scientific literature in the database covers different types of biotic and abiotic stresses with experimental evidence. The Medline resource contains enormous information that could help establish a relationship between biological concepts. For example, CoPub (a literature-based tool) was developed to extract information on genes from humans, mice, and rats and link them to keywords that describe the function of the genes. CoPub was developed based on the assumption that the co-occurrence of a gene and a biomedical concept in the same abstract shows a link between the gene and the concept [17]. Similar to the information stored in our MongoDB collection, we hypothesized that the plant stress genes mentioned with keywords that match GO terms and functional sentences in the same Medline abstract are linked together. The MongoDB collection will be used to complement existing tools to analyze gene expression data in plants during stress. The information from our database will be used to find the function of plant stress genes as they are described by orthologues or the same plant from the scientific articles stored in the MongoDB.

4.4.2 MongoDB database/database content

MongoDB is a non-structural database, and it stores the information in a document-oriented structure. Each document contains a doc field, a text field, objects, and a list of properties in an array . Unlike the SQL database, MongoDB uses collections (collections have no constraint and the fields can contain different data types) instead of data-structured tables [61]. Each document in the MongoDB

collection includes a unique identifier as ID, unique PudMed ID, gene name, the UniProt accession of the gene mention and Entrez ID (if applicable) and GO term) or functional annotation sentence. One issue is due to that GeneRIF (ePMC) annotations are not extensive for plants. Hence not all the functional annotation sentences were retrieved for most of the normalized gene due to the limitation of GeneRIF to the collection of functional annotation sentences.

From the analysis of the dataset, we identified the significant GO terms and the number of publications that mention those GO terms. The GO terms with a frequent mention in the scientific literature collected include; biosynthesis (GO:00009058), metabolism (GO:0008152), defense response (GO:0006952), and photosynthesis (GO:0015979). Similarly, most of the stress-responsive genes in our database mapped to these GO terms. This finding suggests that many stress-responsive genes are involved in biological processes of biosynthesis, metabolism, and defense response. GO, and KEGG enrichment analysis of selected transcriptome studies in plants during stress identified metabolic and biosynthetic processes among significantly enriched terms [21,62]. Using MongoDB query operators, we can insert, update, read, and delete data from MongoDB. The stored documents in MongoDB can easily be retrieved for analysis. In the future, we would like to explore other methods to enable us to enrich the bibliome content related to stress response in plants.

Table 4.1: Statistics of dataset stored in MongoDB

Name (No. of)	**(MongoDB)**
Total abstracts	2,766
Abstracts with functional annotation sentences	1,007
Abstracts with stress genes and GO terms	1,759
Unique plant stress genes	3,716

Unique GO terms	861
Plant species	162

4.4.3 Proof of concept

We used information from the database created in this study to find additional processes of stress-responsive genes in switchgrass. These stress-responsive genes were identified in response to single drought stress and combination of drought and heat stress [19] We identified seven genes with annotation from orthologs that could be further explored and inferred in switchgrass (shown in table S1). The GO terms were retrieved using our pipeline and were not found in the GO database or TAIR. For example, a switchgrass transcript Pavir.1KG530100.v4.1 that encodes TTPS11 gene was upregulated from the switchgrass data during single drought stress at time point DT 168 h. From our MongoDB collection, we found additional information from literature (PMID: 23430324)[63] of upregulation of TTPS11 during biotic stress of *Pytophthora cinnamomic* in *Zea mays*. The following GO term biological processes were retrieved from our pipeline and linked to TPS11 but not reported in GO database and TAIR at the time of collecting the data. They include; defense response (GO:0006852), biosynthesis (GO:0052315), and ethylene biosynthesis (GO:0009693). Similarly, the switchgrass transcript Pavir.5KG421100.v4.1 which encodes WRKY28 gene was identified to be upregulated under single drought stress at time point DT 120 h. From our collection, we found additional processes that WRKY28 could be involved in, which were not reported in TAIR and GO at the time of collecting the data. We found from our database that WRKY28 is involved in oxidative burst (GO:0045730), hyphal growth GO:0030448), and seed germination (GO:0009845).

Interestingly, our database contains information that show an upregulation of WRKY28 under oxalic acid and *Sclerotinia sclerotiorum* stress in Arabidopsis. We may conclude that TPS11 and WRKY28 could be regulated during both biotic and abiotic stress based on the evidence outlined here. Another transcript from the switchgrass RNA-Seq data; Pavir.9NG725100.v4.1 which encode the phosphomannomutase (PMM) gene was upregulated during combined drought and heat stress [19] Our collection revealed that PMM (DoPMM) was also upregulated during cold and salt stress in *Dendrobium officinale* [64]. The same gene was identified in response to oxidative stress in Arabidopsis [65]. The following processes were found to link to PMM gene but were found in GO, TAIR or UniProt at the time of the data collection- biosynthesis (GO:0009058) and seed germination (GO:0009845). Other additional processes linked to the switchgrass drought and heat-responsive genes have been highlighted in supplemental table S1.

From the seven genes switchgrass stress-responsive genes , we selected Pavir.1KG386300.v4.1, which encodes PAL1 as a candidate gene to characterize further and validate. PAL is an essential enzyme in plants due to its role in the phenylpropanoid pathway [29,66]]. Several secondary metabolites such as flavonoids, anthocyanins, and lignin have been derived from phenylpropanoid [32]. Phenylpropanoids play an essential role in plant responses to biotic and abiotic stresses by providing stability against environmental damage [67]. Although PAL has been extensively studied in plants, it is yet to be characterized as an important bioenergy crop in switchgrass.

Four genes encoding PAL were differentially expressed under a combination of drought and heat stress in the switchgrass RNA-Seq data used in this study [19]

These transcripts are Pavir.1KG386300.v4.1, Pavir.1NG356800.v4.1, Pavir.7NG355800.v4.1 and Pavir.1KG386500.v4.1. The transcript Pavir.1KG386300.v4.1 was selected since it showed differential expression at two different time points of imposing drought and heat: DTHT 96/24 h and DTHT 168/96 h. Interestingly all the four transcripts were identified in switchgrass with combined drought and heat but not in switchgrass that only received drought stress [19]. This gene maps to Arabidopsis AT2G37040.1 (AtPAL1). The Arabidopsis Information Resource (TAIR) database listed the following annotation to *AtPAL1*; Cinnamic acid biosynthetic process, defense response, drought recovery, lignin catabolic process, L-phenylalanine catabolic process, pollen development, response to karrikin, response to oxidative stress, response to UV-B and salicylic catabolic process.

Evidence of potential processes of the Pavir.1KG386300.v4.1 transcript that code for PAL1 gene (PavirPAL1) from the MongoDB collection: We found an additional role that PAL1 could play in switchgrass and plants in general. These other annotations were not provided by TAIR or the GO resource Evidence of the part of PAL1 in, cold acclimation (GO:0009631), secondary metabolism (GO:0019748), anthocyanin synthesis (GO:0009718), and biosynthesis (GO:0009058) were retrieved using our pipeline from the literature **PMID: 27439459** (*Arabidopsis thaliana*) [67] and **PMID:26389875** (*Solenostemon scutellarioides*) [29]. We have described the steps to provide the basis to infer these functions (GO terms) to switchgrass PAL1. From the above publications, PAL1 was also found to be induced by cold [68]and light [29]. The above findings complement the role of PAL1 in response to abiotic stresses, which include drought, heat, cold and light stress.

4.4.4 Recombinant PavirPAL1 synthesis

PavirPAL1 gene was synthesized by codon optimization. Codon optimization was necessary for efficient heterologous expression and recombinant protein production [69]. The construct of PavirPAL1 in Pet30a is shown in Fig.S3. Confirmation of the identity of the protein sequence was established by LC-MS analysis. Extensive sequence coverage was achieved at 99% (Fig. S4). The theoretical mass of PavirPAL1 reveal a size of 78.5kDa (Fig. S5).

4.4.5 Sequence analysis

The GenBank accession number of *PavirPAL1* is OL420680. The full-length cDNA sequence (length 3,360 bp) of *PavirPAL1* contains a 2,160 bp open reading frame (ORF) encoding 719 -amino acid proteins. The Neural Network Promoter Prediction Analysis software revealed the putative transcription start site of PavirPAL1 at 191 bp upstream from the start codon (Fig.S6). We identified a possible TATA box located at -31 upstream of the putative transcription start site in *PavirPAL1*. The TATA-box is important and plays a role in the eukaryotic transcription initiation. PLACE and Plant-Care programs were used to predict the core promoter elements. Elements such as TATA box and CAAT box were identified in the 5′-flanking regions of *PavirPAL1*. In addition, MYB binding sites previously identified in other plant PALs were predicted in this region. MYB transcription factors have been reported to play a role in phenylpropanoid biosynthesis [70]. This suggests the possibility of a MYB transcription factor in the regulation of PavirPAL1. Besides, cis-acting elements related to light-responsive elements (GATA-motif) and hormone-responsive elements such as ABRE, MeJA and TGACG-motif were identified in the 5′-flanking regions of *PavirPAL1*.

A multiples sequence alignment showed that the deduced peptide sequence of *PavirPAL1* has high similarity to other known plant PALs, sharing a similarity of 78% identity to *DcPAL* (JQ765748), 76% identity to *SmPAL1* (ABR14606), 77.5% identity to *SsPAL1* (JQ975419), 77% identity to *PcPAL* (CAA57056), 76% identity to *OsPAL* (AK102817), 76% identity to *AtPAL1* (AEC09341), and 72% identity to *ZmPAL1* (AAL40137) (Fig. 6). The deduced peptide sequence contains the active site sequence of phenylalanine and histidine ammonia-lyses: GTVTASGDLVPLSYIAG (position 201-217) (Fig. 4). The active site of all the PAL proteins contains the active site Ala-Ser-Gly (205-207), forming a 3,5 dihydro-5-methylidene-4H-imidazole-4-one (MIO) group.

4.4.6 Characterization of the full-length cDNA sequence of PavirPAL1

The functional domain prediction using the Conserved Domain Database (CDD), Pfam, InterProScan, and SMART [71] predicted that *PavirPAL1* matches to the signature of histidine and phenylalanine ammonia-lyase. It contains the Lyase_aromatic domain (67-583 aa) and belongs to the Lyase class I_family of enzymes that catalyzes beta-elimination reactions and active as homotetramers (**Fig. S1**). The SMART output also showed a compositionally biased region (18-30 aa). Unlike AtPAL1 and SsPAL1, the MotifFinder detected two motifs in the putative PavirPAL1 protein sequence, i.e., Ribosomal L30 domain (accession: PF0707) in the N-terminal region and the aromatic amino acid lyase (accession: PF00221). A ScanProsite of

PavirPAL1 detected the Prosite signature of phenylalanine and histidine ammonia-lyases signature "**GTITASGDLV PLSYIA".** The Prosite scan detected a predicted feature of modified residue of 2,3-didehydroalanine on S212. The secondary

and tertiary structures of *PavirPAL1* were predicted to obtain insight into the structure and function of the gene. The SOMPA program was used to predict the secondary structure of PavirPAL1. The secondary structure showed that α-helices (55.22%) were the main structural components (31.15%) of the random coil (Fig. 5a). The red, green, blue, and pink regions indicate the extended strand, beta-turn, alpha-helix, and random coil, respectively [56]. To better characterize the PavirPAL1 protein, comparative modeling of the three-dimensional (3D) structure of PavirPAL1 was predicted on sequence homology-based using SWISS-MODEL [72]. The 3D structure of PAL from *Petroselinum crispum* [54] have been reported using X-ray crystallography. Similar to the findings of the 3D structures of *P. crispum* and *S. miltiorrhiza*, analysis of PavirPAL1 revealed a "sea horse" shape (Fig.4.5b) as reported for other PALs. Also, SmPAL1 and SsPAL1 structure analysis predicted a similar structure [24]. The PavirPAL1 predicted tertiary structure composed an MIO domain, a core domain, and an inserted shielding domain. In addition, *PavirPAL1* contained the highly conserved Ala-Ser-Gly triad, which served as the MIO prosthetic group site for non-oxidation deamination –([25,29,55]. These conserved sites were identified in *SmPAL1*, *SsPAL1*, and *PcPAL1* which indicates that PavirPAL1 might have the same catalytic activity as other PAL proteins. Our finding was consistent with the description of the X-ray crystallography structures of PcPAL, , and predicted systems of *SmPAL1* and *SsPAL1*.

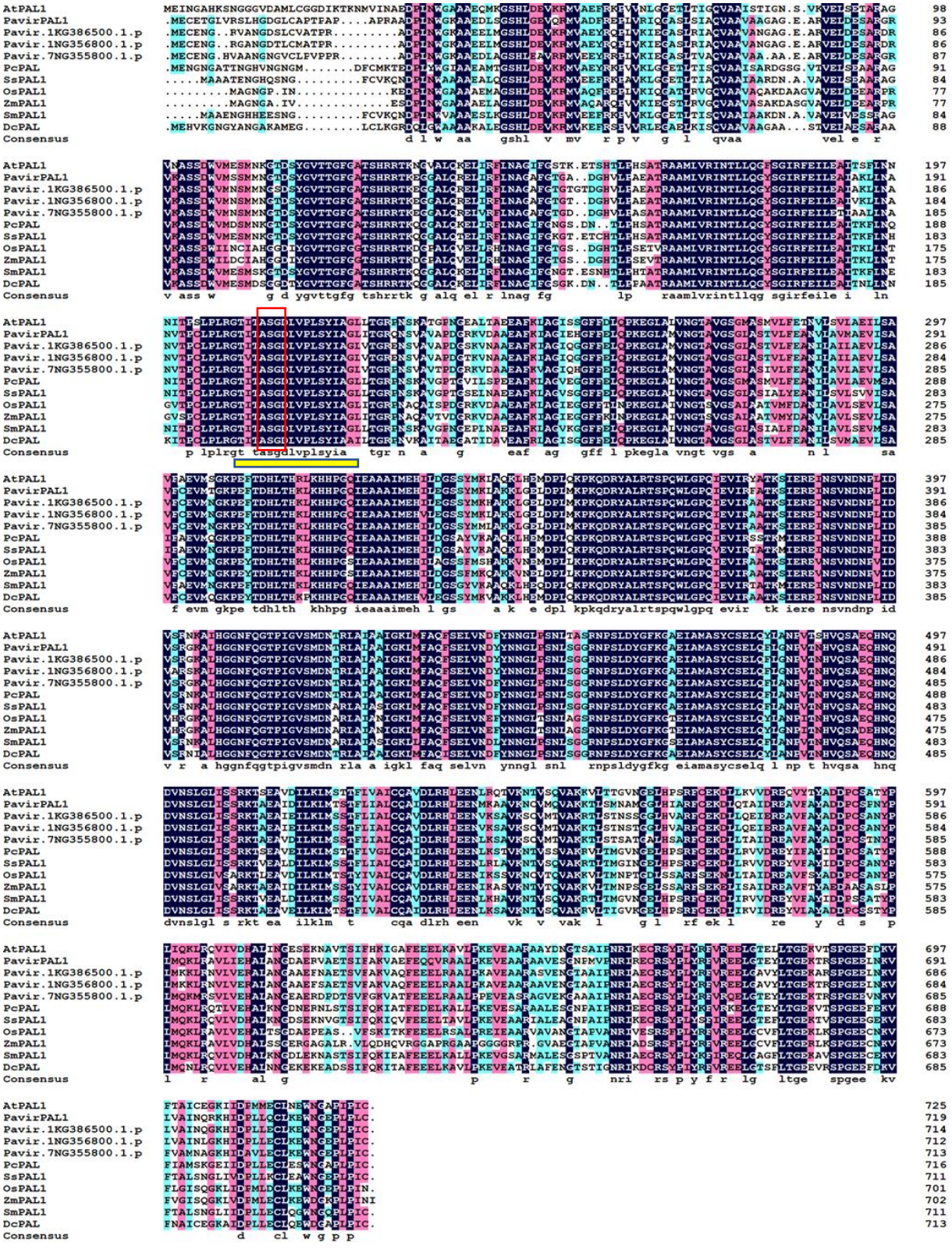
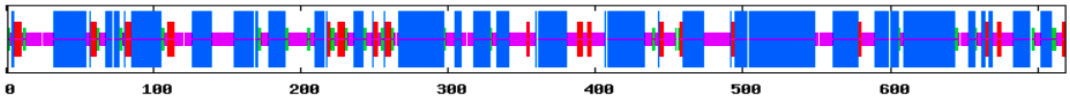


Figure 4.3: Multiple sequence alignment of PavirPAL1 with orthologs. The protein sequences shown here are from Arabidopsis thaliana (AtPAL1, P35510), Solenostemon scutellarioides (SsPAL1, L0BXX7), Oryza sativa (OsPAL1, P14717), Zea mays (ZmPAL1, Q8VXG7), Salvia miltiorrhiza (SmPAL1, A9X1W5), putative PavirPAL1 (Pavir.1KG386300.v4.1), other switchgrass PAL1 (Pavir.1NG356800.v4.1, Pavir.7NG355800.v4.1, Pavir.1KG386500.v4.1.) , PcPAL (P24481) Dendrobium candidum (DcPAL, L7SSS6). The highly-conserved active site motif (Ala-Ser-Gly) which can be converted into a MIO prosthetic group (Zhu et al. 2015, Song et al. 2009) is highlighted in a red open box. The conserved PAL protein finger motif is underlined in yellow.

```
10         20         30         40         50         60
MECETGLVRS LHGDGLCAPT PAPAPRAADP LNWGKAAEDL SGSHLGEVQR MVADFREPLV

        70         80         90        100        110        120
RIQGASLSIA QVAAVAAGAG EARVELDESA RGRVKASSDW VMSSMMNGTD SYGVTTGFGA

       130        140        150        160        170        180
TSHRRTKEGG ALQRELIRFL NAGAFGTGAD GHVLPAEATR AAMLVRINTL LQGYSGIRFE

       190        200        210        220        230        240
ILEAIAKLLN ANVTPCLPLR GTITASGDLV PLSYIAGLIT GRQNSVAVAP DGRKVDAAEA

       250        260        270        280        290        300
FKIAGIEHGF FELQPKEGLA MVNGTAVGSG LASTVLFEAN VLAVMAEVIS AVFCEVMTGK

       310        320        330        340        350        360
PEFTDHLTHK LKHHPGQIEA AAIMEHILEG SSYMKLAKKL GELDPLMKPK QDRYALRTSP

       370        380        390        400        410        420
QWLGPQIEVI RFATKSIERE INSVNDNPLI DVSRGKALHG GNFQGTPIGV SMDNTRLALA

       430        440        450        460        470        480
AIGKLMFAQF SELVNDYYNN GLPSNLSGGR NPSLDYGFKG AEIAMASYCS ELQFLGNPVT

       490        500        510        520        530        540
NHVQSAEQHN QDVNSLGLIS SRKTAEAIDI LKLMTSTFLI ALCQAIDLRH LEENMKAAVK

       550        560        570        580        590        600
NCVMQVAKKT LSMNAMGGLH IARFCEKDLQ TAIDREAVFA YADDPCSPNY PLMQKLRAVL

       610        620        630        640        650        660
IEHALANGDA ERVAETSIFA KVAEFEQQVR AALPKEVEAA RAAVESGNPM VPNRIRECRS

       670        680        690        700        710
YPLYRFVREE LGTEYLTGEK TRSPGEELNK VLVAINQRKH IDPLLQCLKE WNGEPLPLC
```

Figure 4.4: Amino acid sequence of PavirPAL1; the phenylalanine and histidine ammonia-lyases signature(GTITASGDLVPLSYIA) are highlighted in bold. The deamination sites (L-209, V-210, L-259, A-260) are underlined and the catalytic active sites (N-263, G-264, NDN:385-387 aa, HNQDV: 489-493 aa) are indicated with black dots.**(a)**

A)



B)

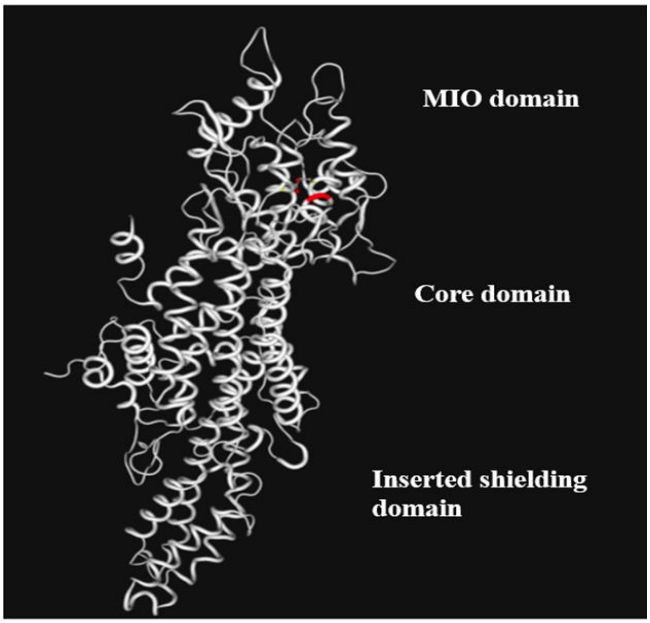


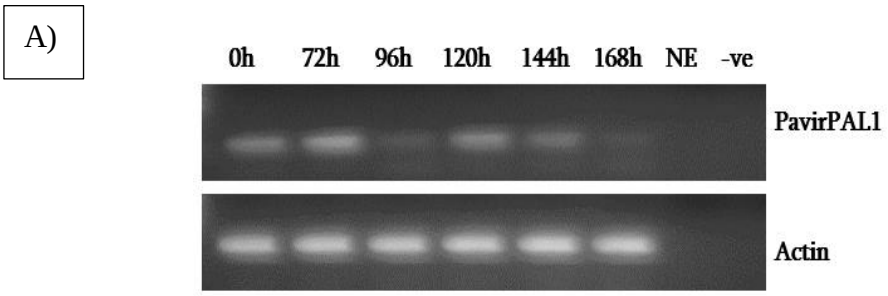
Figure 4.5: Structure analysis. a) Predicted secondary structure of PavirPAL1. The red, green, blue and pink regions represent the extended strand, beta turn, alpha helix and random coil respectively. b) Predicted tertiary structure of PavirPAL1 developed by homology-based modeling. The Ala-Ser-Gly MIO ring is marked in red.

4.4.7 Transcription profile of *PavirPAL1*

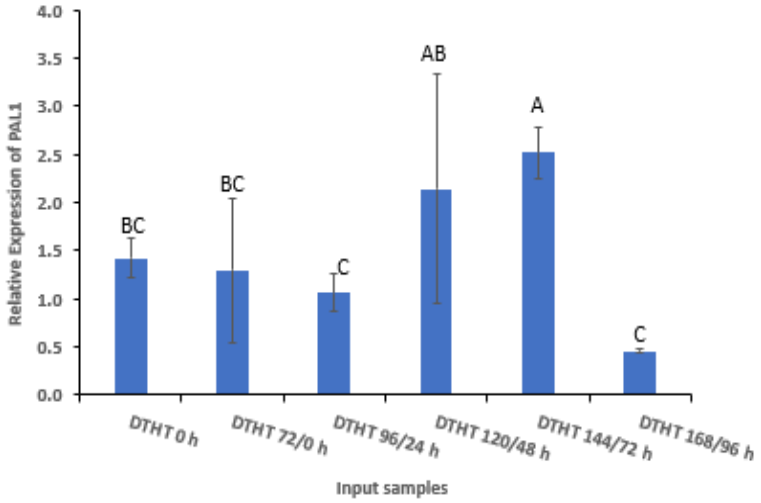
Various studies have shown that the expression of PAL is induced by environmental factors such as pathogen infection, drought, wounding, UV irradiation, and cold temperatures [29,67]. To validate the expression of PavirPAL1 from switchgrass RNA-Seq data during drought and heat stress, RNA samples from the same plant material were used to synthesize cDNA and performed qPCR. As indicated in Figure 9, qPCR results indicate that the expression of *PavirPAL1* was

responsive to a combination of drought and heat stress compared to a single drought stress treatment [19]. The expression of *PavirPAL1* was down-regulated with combined drought and heat stress at time point DTHT 96/24 h and increased at time points DTHT 120/48 h and DTHT 144/72 h. With prolonged exposure to extreme drought and heat stress, *PavirPAL1* reduced markedly at DTHT 168/96 h.

In contrast to our finding, the expression of *SmPAL1* was induced within a short time of drought stress and increased significantly within 30 minutes, followed by a reduction [25]. PAL genes have been induced by various stresses such as light, drought, mechanical wounding, low temperature, UV irradiation, and other stresses [25,29]. The transcription level of PAL from coleus was reduced by dark and under UV-B and wounding treatments. The expression level of PAL has mostly depended on the type of stress imposed and the plant species. For example, the expression of AtPAL1 decreased significantly with ABA treatment; however, the expression of PAL1 from *Salvia miltiorrhiza* (*SmPAL1*) distinctly increased with 100 µM ABA treatment [25,73].



B)


Relative Expression of PAL1
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
BC
BC
C
AB
A
C
DTHT 0 h
DTHT 72/0 h
DTHT 96/24 h
DTHT 120/48 h
DTHT 144/72 h
DTHT 168/96 h
Input samples

C)

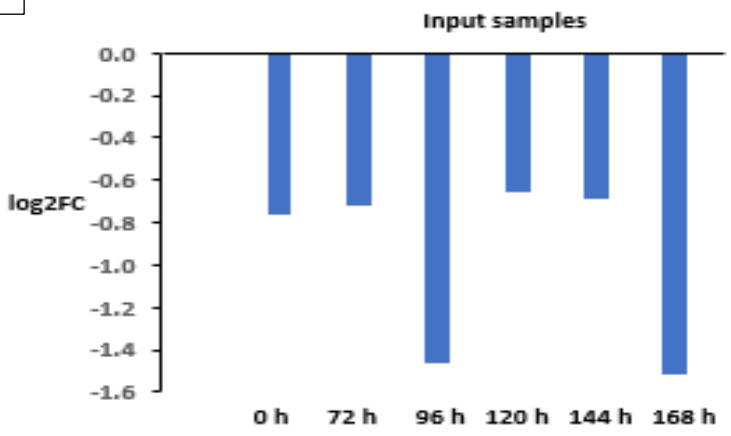

Input samples
log2FC
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
-1.4
-1.6
0 h
72 h
96 h
120 h
144 h
168 h

Figure 4.6: . Expression analysis of PAL1 using leaf tissues from switchgrass at different time points during combined drought and heat stress. (a) Traditional PCR was conducted using PAL1 primers from switchgrass. The primers were designed from the transcripts of PAL1 from switchgrass (Pavir.1KG386300.1). RNA was isolated from the same samples used for the RNA-Seq analysis and cDNA synthesized. Negative controls used in the PCR include NE (no reverse transcriptase enzyme) from the cDNA synthesis and water which is indicated as "-ve". (b) validation of PavirPAL1by qPCR. Three biological replicates and two technical replicates were used for the analysis. Data was analyzed using ANOVA of Minitab statistical software. The different alphabets in the figure indicate statistically significant (p-values<0.05) difference in relative expression of PavirPAL1 between time points (c) The log2FC of Pavir.1KG386300.1 from the switchgrass RNA-Seq data during drought and combination of drought and heat stress.

4.4.8 Expression and purification of recombinant PavirPAL1 in *E.coli*

Until now, many PAL genes from plants have been cloned and expressed in vitro. For example, PAL from *Zea mays* [28], *Arabidopsis thaliana* [26], *Solenostemon scutellarioides* [29], *Juglans regia* [33], *Bambusa oldhamii* [32], and *Petroselium crispum* [73] have been successfully expressed invitro. To confirm the function of PavirPAL1, the recombinant PavirPAL1 was expressed in E.coli BL21(DE3) and purified using TALON Metal affinity Resin+Superdex 200. SDS-PAGE gel showed the various fractions from purification (Fig. 10). This finding indicates that PavirPAL1 was successfully expressed and purified in *E. coli*. *PavirPAL1* expressed a recombinant protein whose molecular weight was about ~78.50 kDa (lane 3 of Western blot image of Fig. 10), which agreed with the predicted mass of 77.68 kDa by the ProtParam online tool. LC-MS confirmed the mass of PavirPAL1. Also, the time course for the expression of the recombinant PavirPAL1 was examined. Our findings indicate that the maximal level of the protein expression

was achieved at 15° C for 16 h after IPTG induction and detected at a solubility of 40 %.

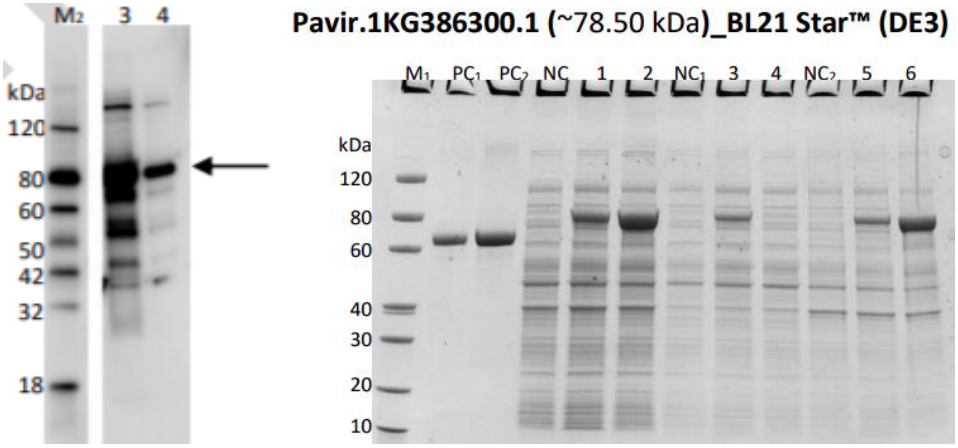


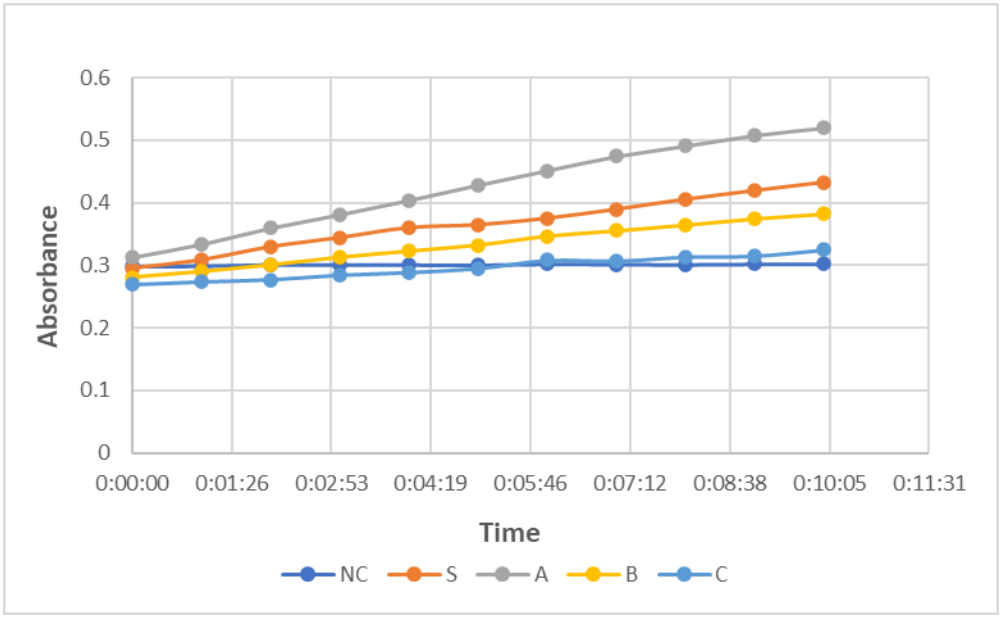
Figure 4.7: Expression and purification of recombinant PavirPAL1 protein in E. coli strain BL21. SDS-PAGE (right) and Western blot (left, using anti-His antibody (GenScript, Cat. No. A00186) analysis of Pavir.1KG386300.1 in E.coli expression construct pET-30a(+). Lane M1: Protein marker Lane M2: Western blot marker Lane PC1: BSA (1 µg) Lane PC2: BSA (2 µg) Lane NC: Cell lysate without induction Lane 1: Cell lysate with induction for 16 h at 15 ℃ Lane 2: Cell lysate with induction for 4 h at 37 °C Lane NC1: Supernatant of cell lysate without induction Lane 3: Supernatant of cell lysate with induction for 16 h at 15 ℃ Lane 4: Supernatant of cell lysate with induction for 4 h at 37 °C Lane NC2: Pellet of cell lysate without induction Lane 5: Pellet of cell lysate with induction for 16 h at 15 ℃ Lane 6: Pellet of cell lysate with induction for 4 h at 37 ° C.

4.4.9 Biochemical characterization of PavirPAL1

The recombinant PavirPAL1 protein purified from switchgrass was analyzed for PAL activity with different concentrations of the test sample (*PavirPAL1*) and standard sample (Phenylalanine ammonia-lyase from Rhodotorula glutinis, Sigma Cat. No. P1016). The result showed that the activity of recombinant PavirPAL1 increased

steadily with increasing concentration and time at a temperature of 30 ° C and wavelength of 270 nm (Fig. 11). The activity of recombinant was higher than that of *R. glutinis*. A linear equation for the change in absorbance of samples of different concentrations within 10 minutes was calculated (calculation in table S3) to determine the absorbance value of the standard sample. The change in absorbance value of the standard sample within 10 minutes was used to calculate the activity of the test samples. The analysis showed a correlation between the concentration of the samples and the absorbance (Fig. 11b). The enzyme activity was determined for three different concentrations of the test samples by measuring the absorbance of the formation of cinnamic acid. The average enzyme activity of PavirPAL1 was calculated, which was about 2.62 U/mg.

A)



B)

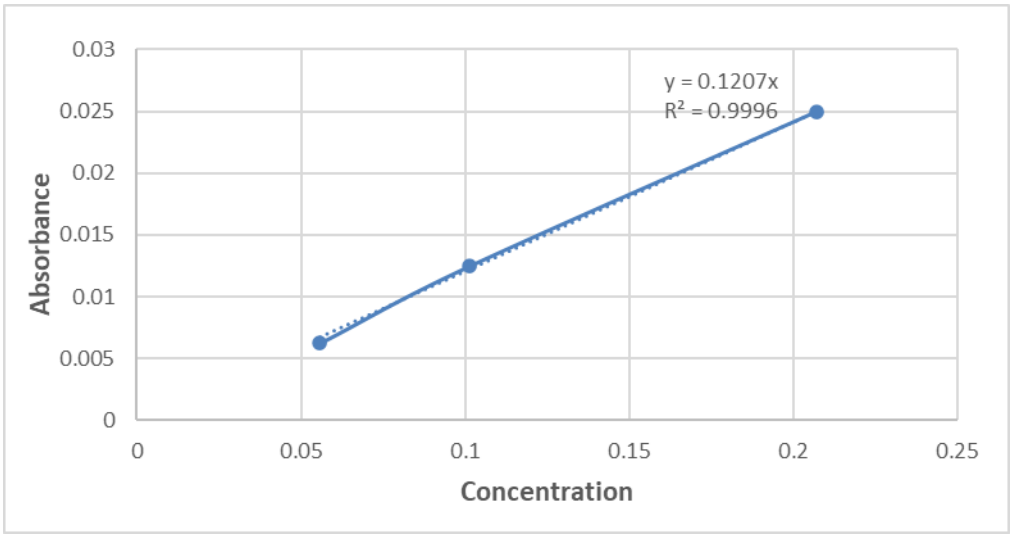


Figure 4.8: Biochemical characterization of purified PavirPAL1. a) calculate change in absorbance value for different concentrations of samples, NC (negative control), S (standard sample-R. glutinis), A, B, C (0.025 mg/ml, 0.0125 mg/ml, 0.00625 mg/ml of test concentrations respectively). b) using concentration and change in absorbance value of samples to calculate the activity of PavirPAL1.

The enzyme activity of PAL proteins has been previously reported [28,29,32] Similar to our finding, the enzyme activity of PavirPAL1 (2.62 U/mg) was slightly higher than the activity of PAL1 from *Bambusa oldhamii*, BoPAL1 (2.26 U/mg). This confirms that PavirPAL1 is suitable to catalyze the deamination of L-phenylalanine to trans-cinnamic acid. Although the optimal temperatures for some recombinant PAL activities have been higher, the optimal temperature for PavirPAL1 activity falls within the ranges for AtPALs [26]. The results explain the metabolic network of phenylpropanoid metabolism in switchgrass.

4.5 Conclusion

This study developed a pipeline to retrieve plant stress genes and annotations from the scientific literature. The information was stored in MongoDB and used to predict the function of plant stress genes by homology. As a used case, information from the MongoDB collection was used to provide annotation to a gene that codes for Phenylalanine ammonia-lyase 1 in switchgrass; PavirPAL1. Bioinformatics analysis was performed to establish the similarity of PavirPAL1 to other plant PALs. Multiple sequence alignment and structure analysis revealed highly conserved regions and sequence and structural similarity with functional plant PAL proteins. After codon optimization, the putative PavirPAL1 (OL420680) was successfully expressed in *E. coli*. The recombinant PavirPAL1 showed PAL activity to convert L-phenylalanine to trans cinnamic acid. The expression of PavirPAL1 in response to a combination of drought and heat stress was validated by qPCR. Taken together, our results show that PavirPAL1 is a functional gene, and the annotations (cold acclimation, secondary metabolism, anthocyanin synthesis, and biosynthesis) identified using our pipeline can be inferred to PavirPAL1. In the future, we would like to extend this pipeline in

iTextMine to enable integration with relation extraction tools that may help highlight interesting aspects of the underlying biology.

4.6 Abbreviation

DT : Drought
DTHT: Combined drought and heat stress
PAL: Phenylalanine ammonia-lyase
qPCR: Quantitative real-time PCR
LC-MS: Liquid chromatography-mass spectrometry

4.7 Supporting/ supplementary information

Additonal supporting files for this chapter can be found using this link; (https://drive.google.com/drive/folders/1drk04OFPms9GzLis1eSOnKsxqjwrHLcu?usp=sharing)

4.8 Acknowledgements

The authors acknowledge the assistance of Julie Cowart at the Center for Bioinformatics and Computational Biology at the University of Delaware for her support in data pre-processing.

REFERENCES

[1] Boyer JS. Plant productivity and environment. Science (80-) 1982. https://doi.org/10.1126/science.218.4571.443.

[2] Raza A, Razzaq A, Mehmood SS, Zou X, Zhang X, Lv Y, et al. Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. Plants 2019. https://doi.org/10.3390/plants8020034.

[3] Naika M, Shameer K, Mathew OK, Gowda R, Sowdhamini R. STIFDB2: An updated version of plant stress-responsive transcription factor database with additional stress signals, stress-responsive transcription factor binding sites and stress-responsive genes in arabidopsis and rice. Plant Cell Physiol 2013;54:1–15. https://doi.org/10.1093/pcp/pcs185.

[4] Azodi CB, Lloyd JP, Shiu S-H. The cis-regulatory codes of response to combined heat and drought stress in Arabidopsis thaliana. NAR Genomics Bioinforma 2020;2:1–16. https://doi.org/10.1093/nargab/lqaa049.

[5] Kumar AA, Mishra P, Kumari K, Panigrahi KCS. Environmental stress influencing plant development and flowering. Front Biosci - Sch 2012. https://doi.org/10.2741/s333.

[6] Li X, Li M, Zhou B, Yang Y, Wei Q, Zhang J. Transcriptome analysis provides insights into the stress response crosstalk in apple (Malus × domestica) subjected to drought, cold and high salinity. Sci Rep 2019. https://doi.org/10.1038/s41598-019-45266-0.

[7] Lamers J, Der Meer T Van, Testerink C. How plants sense and respond to stressful environments. Plant Physiol 2020. https://doi.org/10.1104/PP.19.01464.

[8] Borkotoky S, Saravanan V, Jaiswal A, Das B, Selvaraj S, Murali A, et al. The arabidopsis stress responsive gene database. Int J Plant Genomics 2013;2013:3–6. https://doi.org/10.1155/2013/949564.

[9] Prabha R, Ghosh I, Singh DP. Plant Stress Gene Database : A Collection of Plant Genes Responding to Stress Condition. J Sci Technol 2012;1:28–31.

[10] Mousavi_et_al_2009.pdf n.d.

[11] Priya P, Jain M. RiceSRTFDB: A database of rice transcription factors containing comprehensive expression, cis-regulatory element and mutant

information to facilitate gene function analysis. Database 2013;2013:1–7. https://doi.org/10.1093/database/bat027.

[12] Alter S, Bader KC, Spannagl M, Wang Y, Bauer E, Schön CC, et al. DroughtDB: An expert-curated compilation of plant drought stress genes and their homologs in nine species. Database 2015;2015:1–7. https://doi.org/10.1093/database/bav046.

[13] Krallinger M, Valencia A, Hirschman L. Linking genes to literature: Text mining, information extraction, and retrieval applications for biology. Genome Biol 2008. https://doi.org/10.1186/gb-2008-9-s2-s8.

[14] Ding R, Arighi CN, Lee JY, Wu CH, Vijay-Shanker K. pGenN, a gene normalization tool for plant genes and proteins in scientific literature. PLoS One 2015;10:1–23. https://doi.org/10.1371/journal.pone.0135305.

[15] Ferguson C, Araújo D, Faulk L, Gou Y, Hamelers A, Huang Z, et al. Europe PMC in 2020. Nucleic Acids Res 2021;49:D1507–14. https://doi.org/10.1093/nar/gkaa994.

[16] Alako BTF, Veldhoven A, van Baal S, Jelier R, Verhoeven S, Rullmann T, et al. CoPub Mapper: Mining MEDLINE based on search term co-publication. BMC Bioinformatics 2005. https://doi.org/10.1186/1471-2105-6-51.

[17] Frijters R, Heupers B, van Beek P, Bouwhuis M, van Schaik R, de Vlieg J, et al. CoPub: a literature-based keyword enrichment tool for microarray data analysis. Nucleic Acids Res 2008;36:406–10. https://doi.org/10.1093/nar/gkn215.

[18] Junge A, Jensen LJ. CoCoScore: Context-aware co-occurrence scoring for text mining applications using distant supervision. Bioinformatics 2020;36:264–71. https://doi.org/10.1093/bioinformatics/btz490.

[19] Hayford RK, Serba DD, Xie S, Ayyappan V, Thimmapuram J, Saha MC, et al. Global analysis of switchgrass (Panicum virgatum L.) transcriptomes in response to interactive effects of drought and heat stresses. BMC Plant Biol 2022;22:1–23. https://doi.org/10.1186/s12870-022-03477-0.

[20] Ayyappan V, Saha MC, Thimmapuram J. Comparative transcriptome profiling of upland (VS16) and lowland (AP13) ecotypes of switchgrass. Plant Cell Rep 2017;36:129–50. https://doi.org/10.1007/s00299-016-2065-0.

[21] Chen P, Chen J, Sun M, Yan H, Feng G, Wu B, et al. Comparative transcriptome study of switchgrass (Panicum virgatum L.) homologous autopolyploid and its parental amphidiploid responding to consistent drought stress. Biotechnol Biofuels 2020;13:1–18. https://doi.org/10.1186/s13068-020-01810-z.

[22] Li YF, Wang Y, Tang Y, Kakani VG, Mahalingam R. Transcriptome analysis of heat stress response in switchgrass (Panicum virgatum L.). BMC Plant Biol

2013. https://doi.org/10.1186/1471-2229-13-153.

[23] Lovell JT, MacQueen AH, Mamidi S, Bonnette J, Jenkins J, Napier JD, et al. Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. Nature 2021;590:438–44. https://doi.org/10.1038/s41586-020-03127-1.

[24] Morrow WR, Gopal A, Fitts G, Lewis S, Dale L, Masanet E. Feedstock loss from drought is a major economic risk for biofuel producers. Biomass and Bioenergy 2014;69:135–43. https://doi.org/10.1016/j.biombioe.2014.05.006.

[25] Song J, Wang Z. Molecular cloning, expression and characterization of a phenylalanine ammonia-lyase gene (SmPAL1) from Salvia miltiorrhiza. Mol Biol Rep 2009;36:939–52. https://doi.org/10.1007/s11033-008-9266-8.

[26] Cochrane FC, Davin LB, Lewis NG. The Arabidopsis phenylalanine ammonia lyase gene family: Kinetic characterization of the four PAL isoforms. Phytochemistry 2004;65:1557–64. https://doi.org/10.1016/j.phytochem.2004.05.006.

[27] MINAMI E -I, OZEKI Y, MATSUOKA M, KOIZUKA N, TANAKA Y. Structure and some characterization of the gene for phenylalanine ammonia-lyase from rice plants. Eur J Biochem 1989;185:19–25. https://doi.org/10.1111/j.1432-1033.1989.tb15075.x.

[28] Zang Y, Jiang T, Cong Y, Zheng Z, Ouyang J. Molecular Characterization of a Recombinant Zea mays Phenylalanine Ammonia-Lyase (ZmPAL2) and Its Application in trans-Cinnamic Acid Production from l-Phenylalanine. Appl Biochem Biotechnol 2015;176:924–37. https://doi.org/10.1007/s12010-015-1620-4.

[29] Zhu Q, Xie X, Lin H, Sui S, Shen R, Yang Z, et al. Isolation and functional characterization of a phenylalanine ammonia-lyase gene (SsPAL1) from coleus (Solenostemon scutellarioides (L.) Codd). Molecules 2015;20:16833–51. https://doi.org/10.3390/molecules200916833.

[30] Whetten RW, Sederoff RR. Phenylalanine ammonia-lyase from loblolly pine: Purification of the enzyme and isolation of complementary DNA clones. Plant Physiol 1992;98:380–6. https://doi.org/10.1104/pp.98.1.380.

[31] Jiang Y, Xia N, Li X, Shen W, Liang L, Wang C, et al. Molecular cloning and characterization of a phenylalanine ammonia-lyase gene (LrPAL) from Lycoris radiata. Mol Biol Rep 2011;38:1935–40. https://doi.org/10.1007/s11033-010-0314-9.

[32] Hsieh LS, Yeh CS, Pan HC, Cheng CY, Yang CC, Lee P Du. Cloning and expression of a phenylalanine ammonia-lyase gene (BoPAL2) from Bambusa oldhamii in Escherichia coli and Pichia pastoris. Protein Expr Purif 2010;71:224–30. https://doi.org/10.1016/j.pep.2010.01.009.

[33] Xu F, Deng G, Cheng S, Zhang W, Huang X, Li L, et al. Molecular cloning,

characterization and expression of the phenylalanine ammonia-lyase gene from Juglans Regia. Molecules 2012;17:7810–23. https://doi.org/10.3390/molecules17077810.

[34] Weitzel C, Petersen M. Enzymes of phenylpropanoid metabolism in the important medicinal plant Melissa officinalis L. Planta 2010;232:731–42. https://doi.org/10.1007/s00425-010-1206-x.

[35] Müller HM, Van Auken KM, Li Y, Sternberg PW. Textpresso Central: A customizable platform for searching, text mining, viewing, and curating biomedical literature. BMC Bioinformatics 2018. https://doi.org/10.1186/s12859-018-2103-8.

[36] Levchenko M, Gou Y, Graef F, Hamelers A, Huang Z, Ide-Smith M, et al. Europe PMC in 2017. Nucleic Acids Res 2018. https://doi.org/10.1093/nar/gkx1005.

[37] Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: The universal protein knowledgebase. Nucleic Acids Res 2017;45:D158–69. https://doi.org/10.1093/nar/gkw1099.

[38] Sayers E. A General Introduction to the E-utilities Usage Guidelines and Requirements Minimizing the Number of Requests 2010:1–9.

[39] Cohen SP, Leach JE. Abiotic and biotic stresses induce a core transcriptome response in rice. Sci Rep 2019. https://doi.org/10.1038/s41598-019-42731-8.

[40] Lane T, Best T, Zembower N, Davitt J, Henry N, Xu Y, et al. The green ash transcriptome and identification of genes responding to abiotic and biotic stresses. BMC Genomics 2016;17:1–16. https://doi.org/10.1186/s12864-016-3052-0.

[41] Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: Gene-centered information at NCBI. Nucleic Acids Res 2011;39:52–7. https://doi.org/10.1093/nar/gkq1237.

[42] https://phytozome-next.jgi.doe.gov/

[43] https://phytozome-next.jgi.doe.gov/blast-search

[44] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7. https://doi.org/10.1038/msb.2011.75.

[45] Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017;33:3387–95. https://doi.org/10.1093/bioinformatics/btx431.

[46] https://web.expasy.org/protparam/

[47] https://www.ncbi.nlm.nih.gov/cdd/

[48] https://www.ncbi.nlm.nih.gov/structure/lexington/lexington.cgi

[49] https://prosite.expasy.org/scanprosite/

[50] https://www.genome.jp/tools/motif/

[51] https://www.ebi.ac.uk/interpro/search/sequecne-search

[52] Higo K, Ugawa Y, Iwamoto M, Higo H. PLACE: A database of plant cis-acting regulatory DNA elements. Nucleic Acids Res 1998;26:358–9. https://doi.org/10.1093/nar/26.1.358.

[53] Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van De Peer Y, et al. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 2002;30:325–7. https://doi.org/10.1093/nar/30.1.325.

[54] https://swiss-model.expasy.org/

[55] Ritter H, Schulz GE. Structural basis for the entrance into the phenylpropanoid metabolism catalyzed by phenylalanine ammonia-lyase. Plant Cell 2004;16:3426–36. https://doi.org/10.1105/tpc.104.025288.

[56] https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html

[57] Moisă ME, Amariei DA, Nagy EZA, Szarvas N, Toșa MI, Paizs C, et al. Fluorescent enzyme-coupled activity assay for phenylalanine ammonia-lyases. Sci Rep 2020;10:1–11. https://doi.org/10.1038/s41598-020-75474-y.

[58] Fritz RR, Hodgins DS, Abell CW. Phenylalanine ammonia-lyase. Induction and purification from yeast and clearance in mammals. J Biol Chem 1976;251:4646–50. https://doi.org/10.1016/s0021-9258(17)33251-9.

[59] Ayyappan V, Saha MC, Thimmapuram J, Sripathi VR, Bhide KP, Fiedler E, et al. Comparative transcriptome profiling of upland (VS16) and lowland (AP13) ecotypes of switchgrass. Plant Cell Rep 2017. https://doi.org/10.1007/s00299-016-2065-0.

[60] Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. Nat Protoc 2008;3:1101–8. https://doi.org/10.1038/nprot.2008.73.

[61] Boicea A, Radulescu F, Agapin LI. MongoDB vs Oracle - Database comparison. Proc - 3rd Int Conf Emerg Intell Data Web Technol EIDWT 2012 2012:330–5. https://doi.org/10.1109/EIDWT.2012.32.

[62] Li Q, Qin Y, Hu X, Li G, Ding H, Xiong X, et al. Transcriptome analysis uncovers the gene expression profile of salt-stressed potato (Solanum tuberosum L.). Sci Rep 2020;10:1–19. https://doi.org/10.1038/s41598-020-62057-0.

[63] Allardyce JA, Rookes JE, Hussain HI, Cahill DM. Transcriptional profiling of

Zea mays roots reveals roles for jasmonic acid and terpenoids in resistance against Phytophthora cinnamomi. Funct Integr Genomics 2013;13:217–28. https://doi.org/10.1007/s10142-013-0314-7.

[64] He C, Zeng S, Teixeira da Silva JA, Yu Z, Tan J, Duan J. Molecular cloning and functional analysis of the phosphomannomutase (PMM) gene from Dendrobium officinale and evidence for the involvement of an abiotic stress response during germination. Protoplasma 2017;254:1693–704. https://doi.org/10.1007/s00709-016-1044-1.

[65] Cho KM, Nguyen HTK, Kim SY, Shin JS, Cho DH, Hong SB, et al. CML10, a variant of calmodulin, modulates ascorbic acid synthesis. New Phytol 2016;209:664–78. https://doi.org/10.1111/nph.13612.

[66] Liu Y, Zhang X, Tran H, Shan L, Kim J, Childs K, et al. Assessment of drought tolerance of 49 switchgrass (Panicum virgatum) genotypes using physiological and morphological parameters. Biotechnol Biofuels 2015;8:1–18. https://doi.org/10.1186/s13068-015-0342-8.

[67] Vogt T. Phenylpropanoid biosynthesis. Mol Plant 2010;3:2–20. https://doi.org/10.1093/mp/ssp106.

[68] van Buer J, Cvetkovic J, Baier M. Cold regulation of plastid ascorbate peroxidases serves as a priming hub controlling ROS signaling in Arabidopsis thaliana. BMC Plant Biol 2016;16:1–20. https://doi.org/10.1186/s12870-016-0856-7.

[69] Fu H, Liang Y, Zhong X, Pan ZL, Huang L, Zhang HL, et al. Codon optimization with deep learning to enhance protein expression. Sci Rep 2020;10:1–9. https://doi.org/10.1038/s41598-020-74091-z.

[70] Xiao R, Zhang C, Guo X, Li H, Lu H. MYB transcription factors and its regulation in secondary cell wall formation and lignin biosynthesis during xylem development. Int J Mol Sci 2021;22. https://doi.org/10.3390/ijms22073560.

[71] http://smart.embl-heidelberg.de/

[72] Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. Bioinformatics 2006;22:195–201. https://doi.org/10.1093/bioinformatics/bti770.

[73] Mohr PG, Cahill DM. Suppression by ABA of salicylic acid and lignin accumulation and the expression of multiple genes, in Arabidopsis infected with Pseudomonas syringae pv. tomato. Funct Integr Genomics 2007;7:181–91. https://doi.org/10.1007/s10142-006-0041-4.

[74] Appert C, Logemann E, Hahlbrock K, Schmid J, Amrhein N. Structural and Catalytic Properties of the Four Phenylalanine Ammonia-Lyase Isoenzymes

from Parsley (Petroselinum Crispum Nym.). Eur J Biochem 1994;225:491–9. https://doi.org/10.1111/j.1432-1033.1994.00491.x.

Chapter 5

VISUALIZATION OF SWITCHGRASS TRANSCRIPTOME DATA DURING DROUGHT AND HEAT STRESS USING CYTOSCAPE

5.1 Abstract

Cytoscape is an open-source, cross-platform bioinformatics program written in Java. It is used to visualize interaction networks and integrate these with expression profiles and other high throughput data sets. ClueGo is a Cytoscape plug-in that visualizes the non-redundant biological terms for large clusters of genes in a functionally grouped network. The ClueGO network is created with kappa statistics, and it reflects the relationships between the terms which are shared based on the similarity of the associated genes. We used ClueGo to visualize the functionally grouped terms and pathways(KEGG/Reactome) to create a network-based analysis of the transcriptome data on switchgrass drought and heat-responsive genes. Unique and overlapping functional networks of GO terms and pathways of drought and combined drought and heat stress were identified. Furthermore, our analysis revealed a possible link between the enriched terms or pathways, thus providing the basis for conducting experiments to explore the detailed regulation of stress-responsive genes.

5.2 Background

Cytoscape is widely used for network-based data integration. Analysis of gene expression dataset is commonly performed to gain insight into the underlying biological processes. The network represents a relationship between aspects of data,

and therefore performing this activity help to discover functional gene interaction and provides the biologist with functional evidence. In addition, the network analysis enables an effective means of data comparison, data interpretation and generation of hypothesis [1] . ClueGO plug-in is easy to use and strongly enhances the biological interpretation of a large data set [2]. One of the functionalities of Cytoscape software is to link the network to databases of annotations [3]. We visualized the transcriptome data on switchgrass when imposed with a single drought and drought and heat stress combinations to explore biological networks. ClueGo integrates Gene Ontology (GO) terms and KEGG/Reactome pathways and creates a functionally organized GO/pathway term network. By clicking the update option of ClueGO automatically downloads the most recent files of GO, KEGG, and Reactome release at any time [2]. We uploaded all the drought and combinations of drought and heat-responsive genes into Cytoscape. We used the ClueGo plug-in to understand functionally grouped gene ontology terms and pathway annotation networks in the data. We identified common and specific transcriptional responses during a single drought and combinations of drought and heat stress.

5.3 Method

To calculate enrichment values for terms, pathways and groups using ClueGO, we used two-sided (enrichment/depletion) test based on the hypergeometric distribution to calculate doubling for two-sided tests to support the functional conservation effect [4,5]. To control the false positives, we used Bonferroni method to correct the p-values [6]. ClueGO creates first a binary gene-term matrix with the selected terms and their associated genes. Based on this matrix, a term–term similarity matrix is calculated using chance corrected kappa statistics to determine the

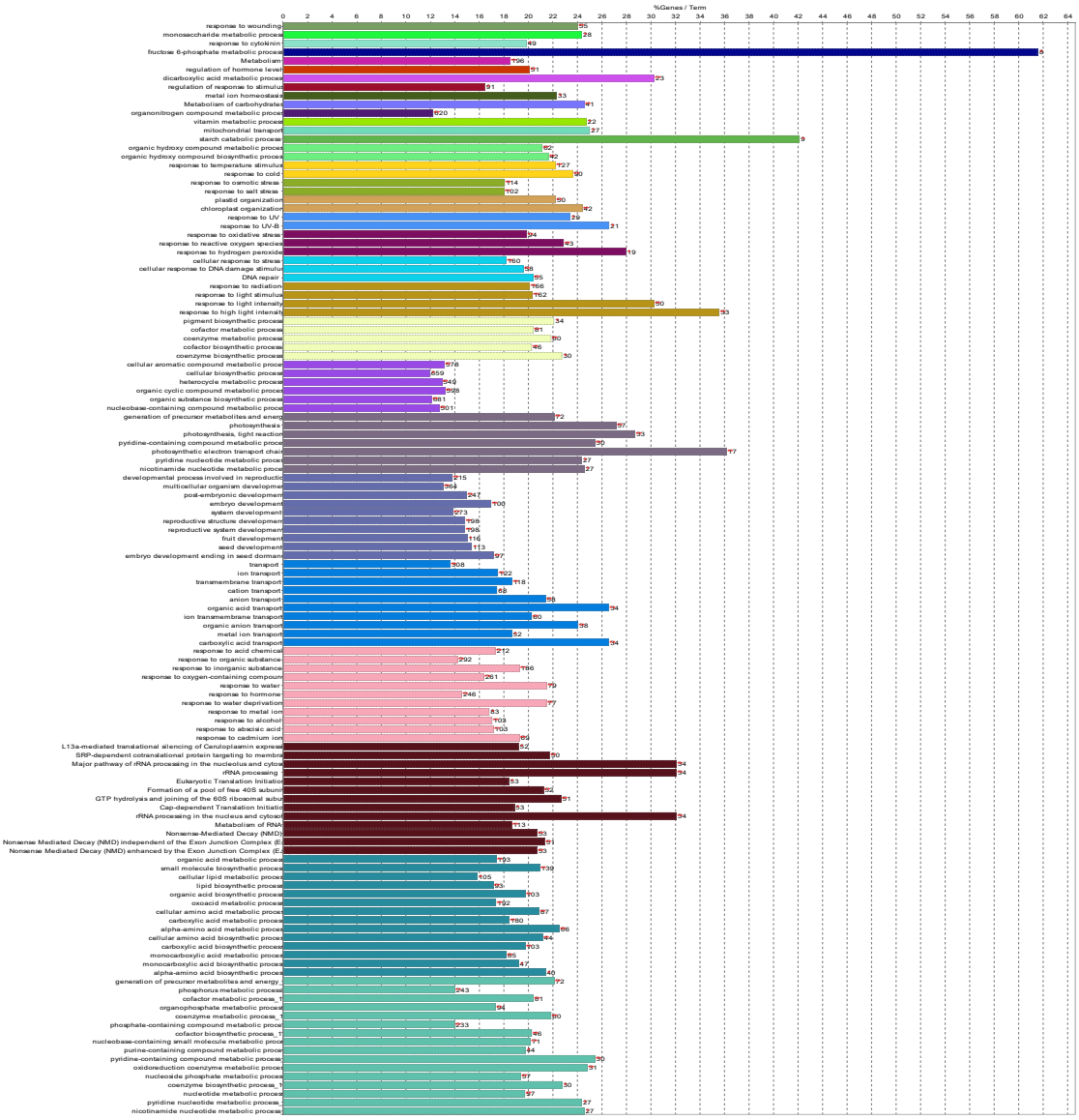
association strength between the terms. Since the term–term matrix is of categorical origin, kappa statistic was found to be the most suitable method. Finally, the created network represents the terms as nodes which are linked based on a predefined kappa score level. The kappa score level threshold can initially be adjusted on a positive scale from 0 to 1 to restrict the network connectivity in a customized way. For this analysis, a kappaScore threshold of 0.3 was used. The size of the nodes reflects the enrichment significance of the term. The functional groups indicated by the most significant terms are visualized in the network, showing their relationships' details. In addition, we included the other ways of selecting the group-leading term such as showing the number of genes per term as shown in **Figure 5.1a**.

5.4 Results and discussion

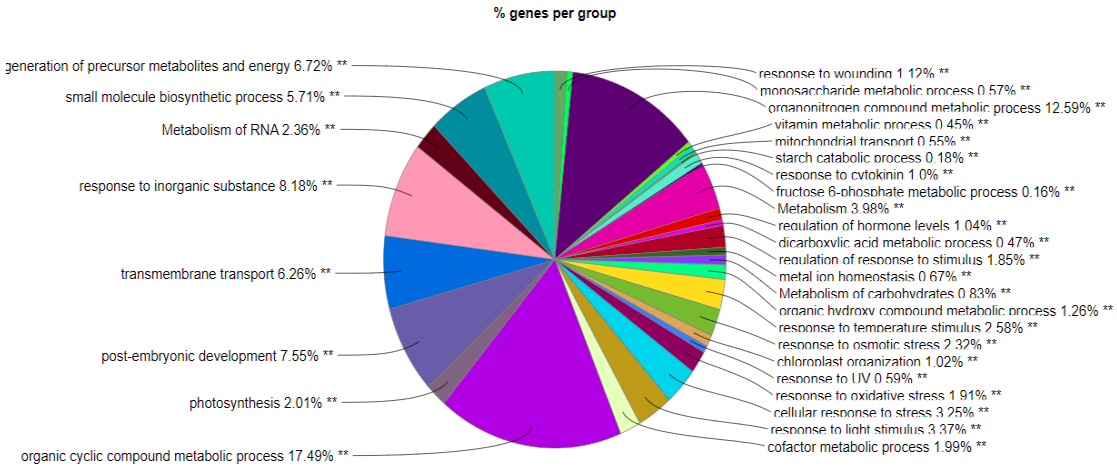
5.4.1 GO enrichment analysis for combined drought and heat (DTHT) differentially expressed genes

Advances in systems biology approaches for integrated functional analysis have contributed to identifying GO-enriched terms, pathways, and networks underlying stress-response mechanisms. This study conducted a network-based analysis of single drought and combined drought and heat differentially expressed genes in switchgrass. Comprehensive analysis of biological pathways in plants under multiple stress situations may be the key step for understanding molecular mechanism underling cross-talk among stress signaling. Visualization of the networks of pathway terms based on the DTHT specific responses reveals the enormous transcriptional responses evoked in switchgrass. ClueGO plug-in of Cytoscape helps to visualize the theme of the pathway or term results in the network.

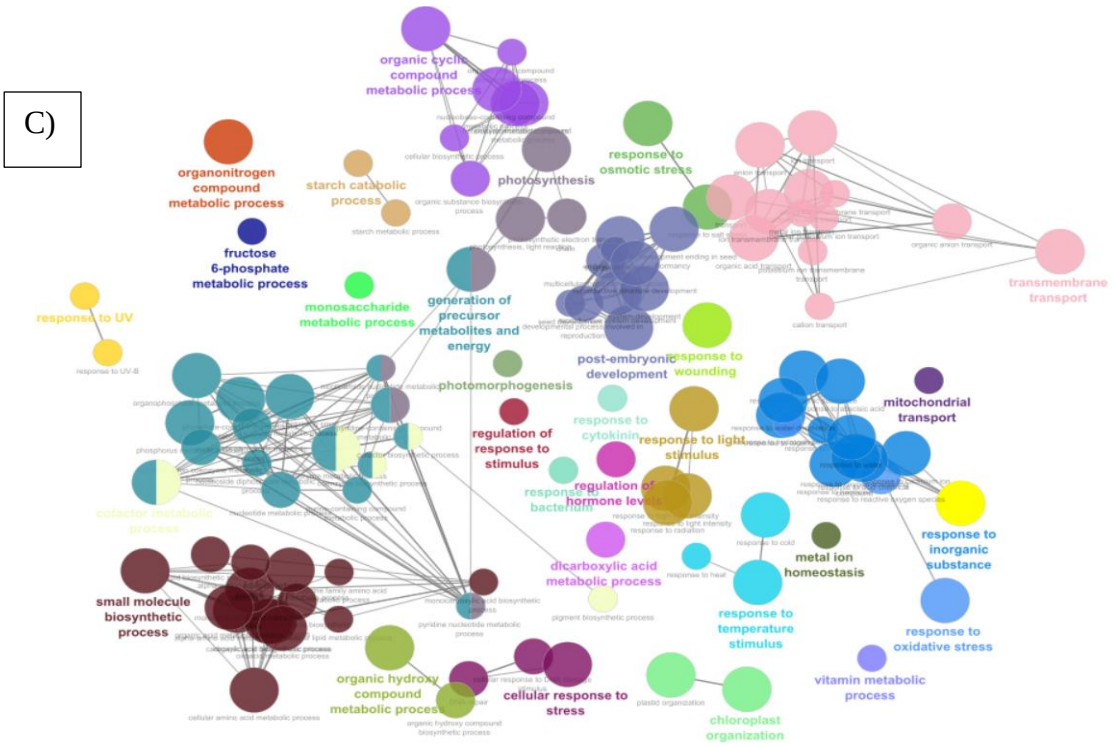
A)


%Genes / Term
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64
response to wounding
monosaccharide metabolic process
response to cytokinin
fructose 6-phosphate metabolic process
Metabolism
regulation of hormone level
dicarboxylic acid metabolic process
regulation of response to stimulus
metal ion homeostasis
Metabolism of carbohydrates
organonitrogen compound metabolic process
vitamin metabolic process
mitochondrial transport
starch catabolic process
organic hydroxy compound metabolic process
organic hydroxy compound biosynthetic process
response to temperature stimulus
response to cold
response to osmotic stress
response to salt stress
plastid organization
chloroplast organization
response to UV
response to UV-B
response to oxidative stress
response to reactive oxygen species
response to hydrogen peroxide
cellular response to stress
cellular response to DNA damage stimulus
DNA repair
response to radiation
response to light stimulus
response to light intensity
response to high light intensity
pigment biosynthetic process
cofactor metabolic process
coenzyme metabolic process
cofactor biosynthetic process
coenzyme biosynthetic process
cellular aromatic compound metabolic process
cellular biosynthetic process
heterocycle metabolic process
organic cyclic compound metabolic process
organic substance biosynthetic process
nucleobase-containing compound metabolic process
generation of precursor metabolites and energy
photosynthesis
photosynthesis, light reaction
pyridine-containing compound metabolic process
photosynthetic electron transport chain
pyridine nucleotide metabolic process
nicotinamide nucleotide metabolic process
developmental process involved in reproduction
multicellular organism development
post-embryonic development
embryo development
system development
reproductive structure development
reproductive system development
fruit development
seed development
embryo development ending in seed dormancy
transport
ion transport
transmembrane transport
cation transport
anion transport
organic acid transport
ion transmembrane transport
organic anion transport
metal ion transport
carboxylic acid transport
response to acid chemical
response to organic substance
response to inorganic substance
response to oxygen-containing compound
response to water
response to hormone
response to water deprivation
response to metal ion
response to alcohol
response to abscisic acid
response to cadmium ion
L13a-mediated translational silencing of Ceruloplasmin expression
SRP-dependent cotranslational protein targeting to membrane
Major pathway of rRNA processing in the nucleolus and cytosol
rRNA processing
Eukaryotic Translation Initiation
Formation of a pool of free 40S subunits
GTP hydrolysis and joining of the 60S ribosomal subunit
Cap-dependent Translation Initiation
rRNA processing in the nucleus and cytosol
Metabolism of RNA
Nonsense-Mediated Decay (NMD)
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)
organic acid metabolic process
small molecule biosynthetic process
cellular lipid metabolic process
lipid biosynthetic process
organic acid biosynthetic process
oxoacid metabolic process
cellular amino acid metabolic process
carboxylic acid metabolic process
alpha-amino acid metabolic process
cellular amino acid biosynthetic process
carboxylic acid biosynthetic process
monocarboxylic acid metabolic process
monocarboxylic acid biosynthetic process
alpha-amino acid biosynthetic process
generation of precursor metabolites and energy
phosphorus metabolic process
cofactor metabolic process
organophosphate metabolic process
coenzyme metabolic process
phosphate-containing compound metabolic process
cofactor biosynthetic process
nucleobase-containing small molecule metabolic process
purine-containing compound metabolic process
pyridine-containing compound metabolic process
oxidoreduction coenzyme metabolic process
nucleoside phosphate metabolic process
coenzyme biosynthetic process
nucleotide metabolic process
pyridine nucleotide metabolic process
nicotinamide nucleotide metabolic process

B)


% genes per group
generation of precursor metabolites and energy 6.72% **
small molecule biosynthetic process 5.71% **
Metabolism of RNA 2.36% **
response to inorganic substance 8.18% **
transmembrane transport 6.26% **
post-embryonic development 7.55% **
photosynthesis 2.01% **
organic cyclic compound metabolic process 17.49% **
response to wounding 1.12% **
monosaccharide metabolic process 0.57% **
organonitrogen compound metabolic process 12.59% **
vitamin metabolic process 0.45% **
mitochondrial transport 0.55% **
starch catabolic process 0.18% **
response to cytokinin 1.0% **
fructose 6-phosphate metabolic process 0.16% **
Metabolism 3.98% **
regulation of hormone levels 1.04% **
dicarboxylic acid metabolic process 0.47% **
regulation of response to stimulus 1.85% **
metal ion homeostasis 0.67% **
Metabolism of carbohydrates 0.83% **
organic hydroxy compound metabolic process 1.26% **
response to temperature stimulus 2.58% **
response to osmotic stress 2.32% **
chloroplast organization 1.02% **
response to UV 0.59% **
response to oxidative stress 1.91% **
cellular response to stress 3.25% **
response to light stimulus 3.37% **
cofactor metabolic process 1.99% **

C)


organic cyclic compound metabolic process
organonitrogen compound metabolic process
starch catabolic process
photosynthesis
response to osmotic stress
fructose 6-phosphate metabolic process
transmembrane transport
response to UV
monosaccharide metabolic process
generation of precursor metabolites and energy
post-embryonic development
response to wounding
photomorphogenesis
mitochondrial transport
response to cytokinin
response to light stimulus
regulation of response to stimulus
regulation of hormone levels
response to bacterium
cofactor metabolic process
response to inorganic substance
dicarboxylic acid metabolic process
metal ion homeostasis
small molecule biosynthetic process
response to temperature stimulus
response to oxidative stress
organic hydroxy compound metabolic process
cellular response to stress
vitamin metabolic process
chloroplast organization

D)

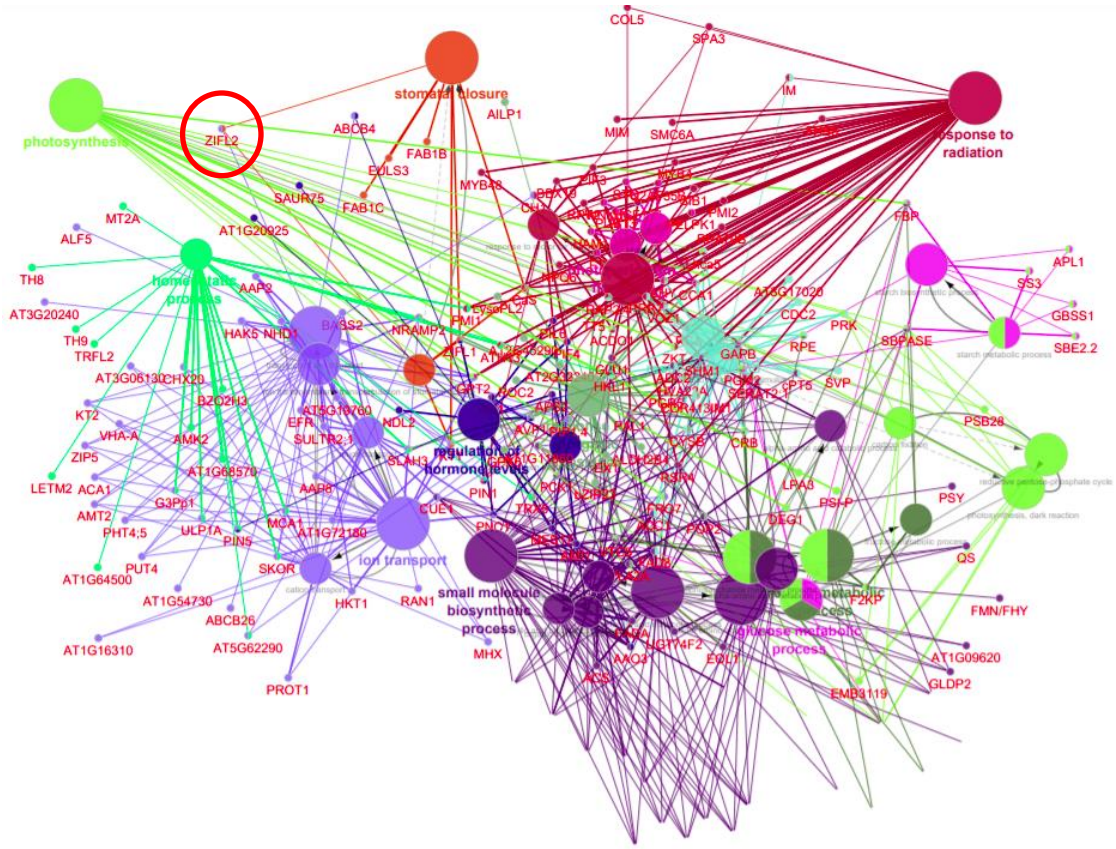


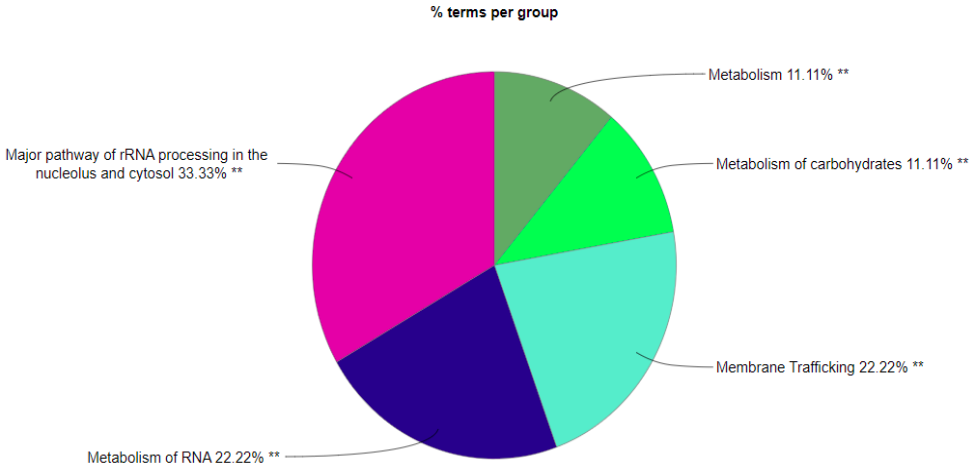
Figure 5.1: Network visualization of enriched terms among the differentially regulated genes during combined drought and heat stress. The network analysis was performed by ClueGo analysis. **a)** GO terms specific for combined drought and heat DEGs from switchgrass. The bars represent the number of genes assigned with the terms. The percentage of genes per term is shown as bar label. **b)** Overview chart with functional groups including specific terms for DTHT DEGs. **c)** Over-represented GO analysis in the DTHT differentially expressed genes. These are functionally grouped network with terms as nodes linked based on their kappa score level (≥0.3), significant terms are shown. The node size represents the term enrichment significance. Functionally related groups partially overlap. The edges are related to the relationships between the selected terms defined based on the genes shared in a similar way. The label of the most significant term is used as the leading group term. **d)** Gene networks for GO biological process terms of DTHT DEGs.

Similar to the enrichment analysis of DTHT DEGs by AgriGO (chapter 2), the ClueGo analysis showed significant enrichment (P-value<=0.05) in these biological processes, organonitrogen compound metabolic process, organic cyclic compound metabolic process, and small molecule biosynthesis process (Figure 5.1 (a,b,c)). The enrichment analysis also highlighted the terms "photosynthesis", "response to oxidative stress", and "response to light stimulus" in the DTHT DEGs. A network-based analysis of the corresponding GO terms (biological processes) of DTHT in genes revealed specific terms enriched by the DTHT DEGs. As indicated in Figure 5.1d. Some significant terms enriched by DTHT genes include 'photosynthesis, 'ion transport', 'stomatal closure', and 'response to radiation'. The genes associated with these terms are shown; for example, the term ‘stomatal closure' reveals genes associated with this term, including FAB1B, and FAB1C. Another gene, ZIFL2 (Zinc-Induced Facilitator 2) was identified with overlapping terms ‘stomatal closure' and 'ion transport' biological processes (highlighted in red in Figure 5.1d). The role of ZIFL2 in potassium and cesium homeostasis has recently been reported. An isoform of ZIFL1, ZIFL1.3, has been reported to play a key role in drought stress by regulating stomatal closure [7]. The presence of ZIFL2 with overlapping terms' ion transport' and 'stomatal closure' reveals the possibility of its role in transport and drought stress by regulating stomatal closure. Visualization of the network of GO terms provides insights to the underlying mechanisms of the stress genes.

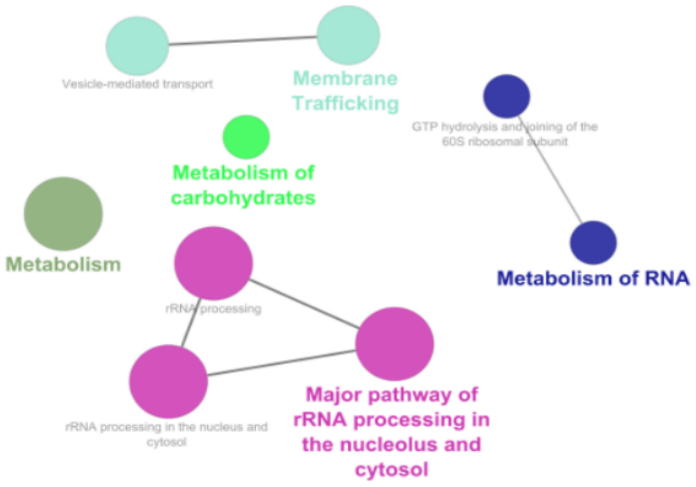
5.4.2 Pathway analysis (Reactome) of DTHT switchgrass genes

A)

B)



C)



D)

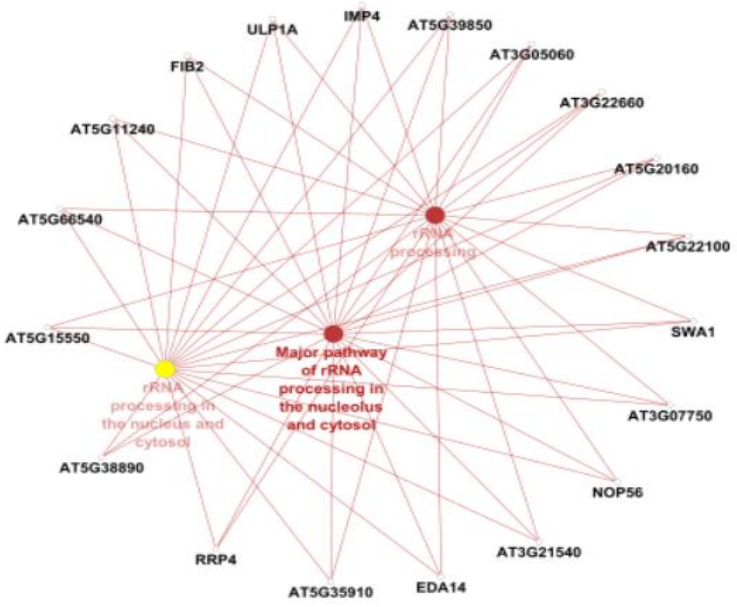
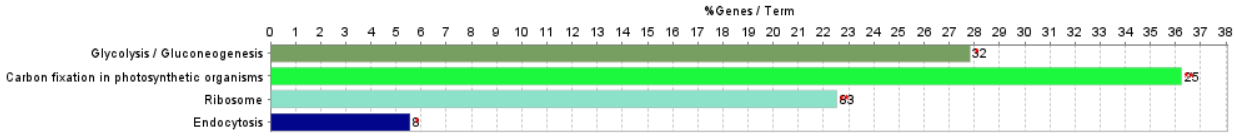


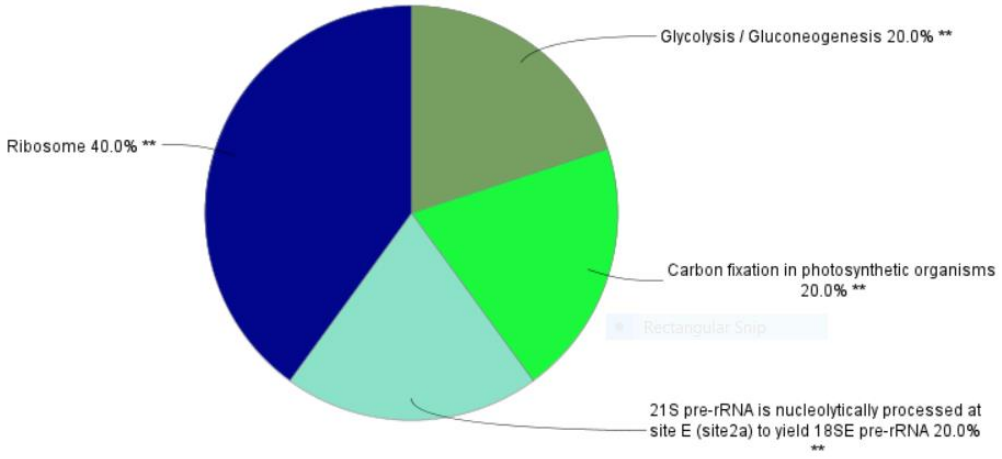
Figure 5.2: **Network visualization of enriched pathways (Reactome) in DTHT gene signature performed by ClueGO analysis.** Pathway analysis (Reactome) of DTHT differentially expressed genes. **a)**The bars represent the number of DTHT genes assigned with the pathways. The percentage of genes per pathway is shown as bar label. . **b)** Overview chart with functional groups including specific pathways for DTHT DEGs. The label of the most significant term is used as the leading group term. **c)** Functionally grouped network with pathways as nodes linked based on their kappa score level (≥0.3), significant pathways for DTHT genes are shown . The node size represents the pathway enriched significance. Functionally related groups partially overlap. The edges are related to the relationships between the selected pathways defined based on the genes shared in a similar way. **d)** Retrieved connection of the common genes of the major pathway enriched by DTHT differentially regulated genes which is "major pathway of rRNA processing in the nucleolus and cytosol"

5.4.3 KEGG Pathway analysis by ClueGO

A)



B)



C)

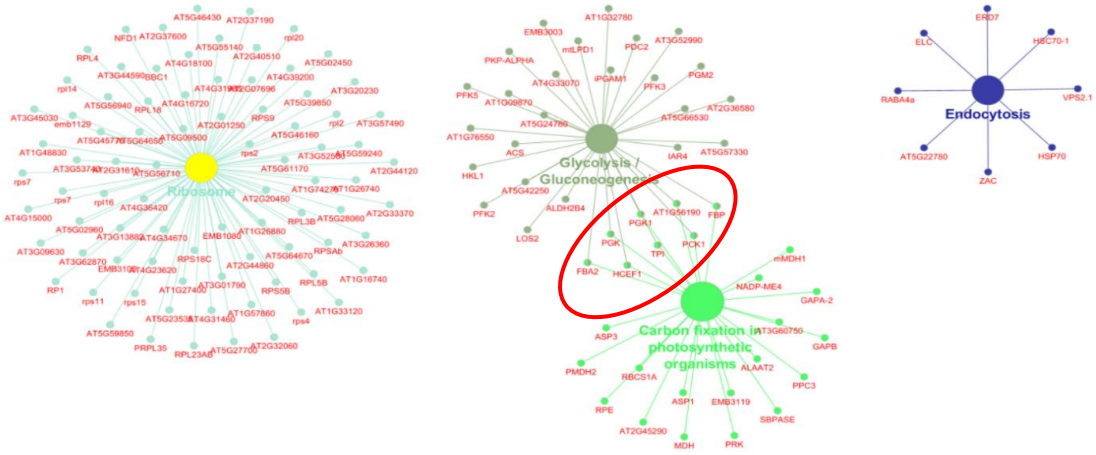
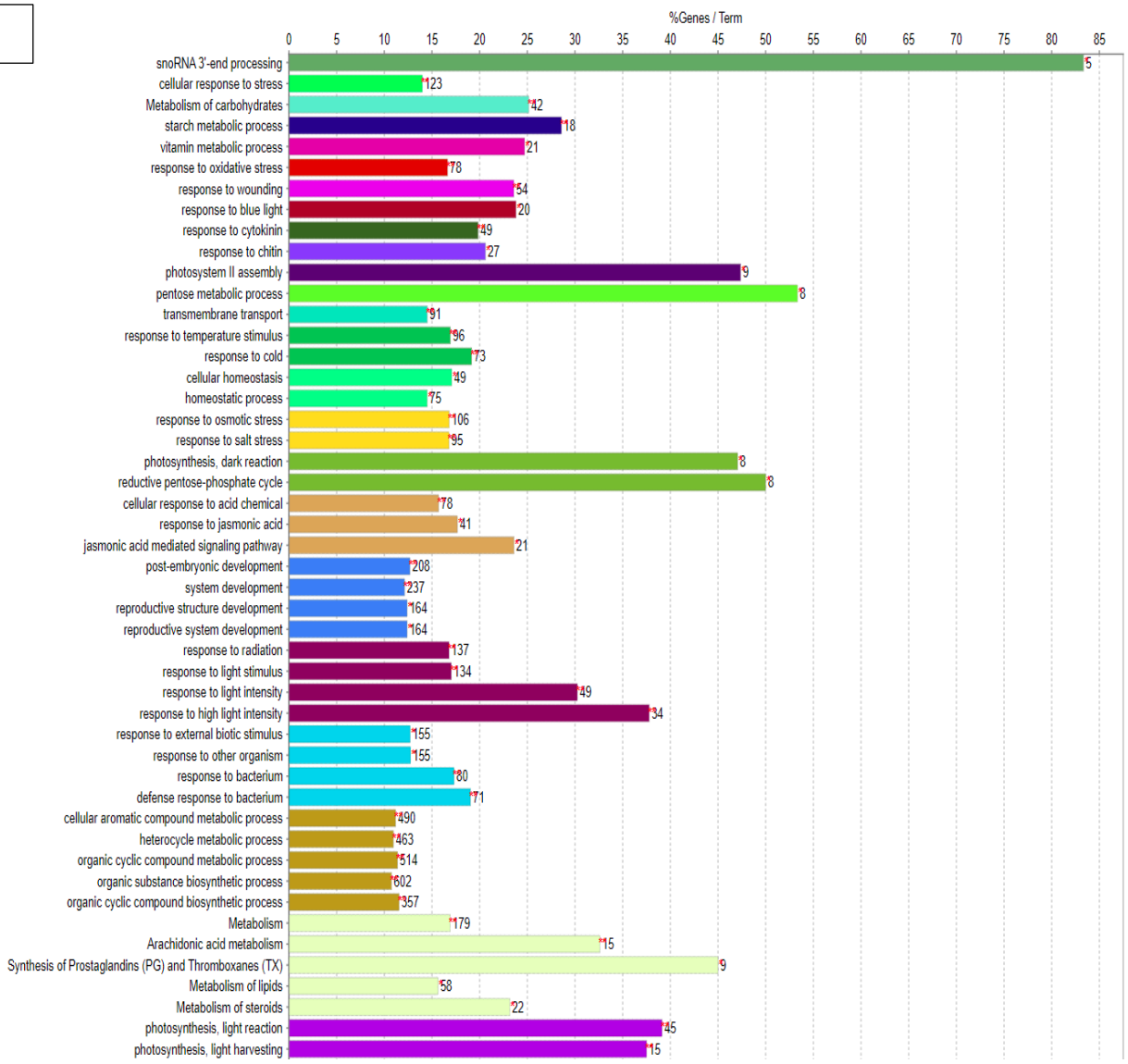


Figure 5.3: **Network visualization of enriched pathways (KEGG) in DTHT gene signature performed by ClueGO analysis.** Pathway analysis (KEGG) of DTHT differentially expressed genes. **a)**The bars represent the number of DTHT genes assigned with the pathways. The percentage of genes per pathway is shown as bar label. **b)** Overview chart of specific pathways for DTHT DEGs. The label of the most significant term is used as the leading group term. **c)** Retrieved connection of the common genes of the significant pathway enriched by DTHT differentially regulated genes. Unique and shared genes between the pathways "glycolysis/gluconeogenesis" and "carbon fixation in photosynthetic organisms" have been circled in red.

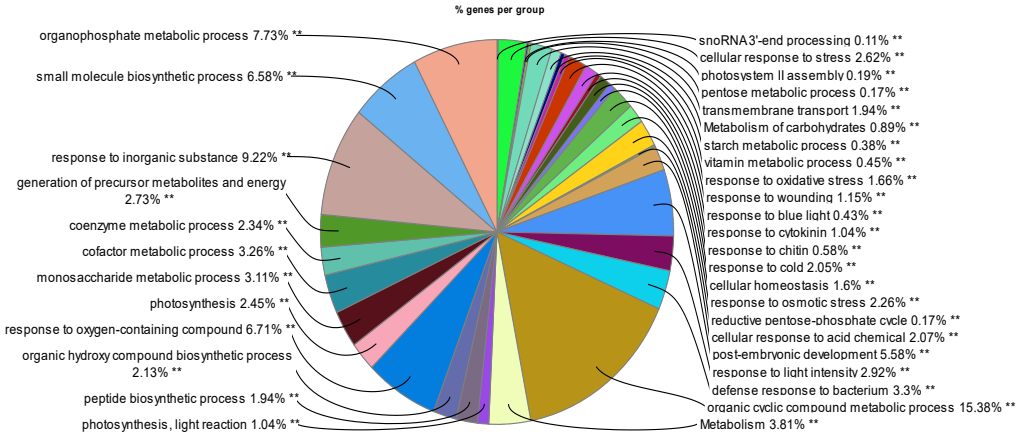
5.4.4 GO enrichment analysis for single drought differentially expressed genes (DEGs)

The enrichment values for terms and groups from ClueGO results of single drought switchgrass genes are presented. The functional groups represented by their most significant (leading) term is visualized in the network providing an insightful view of their interrelationships.

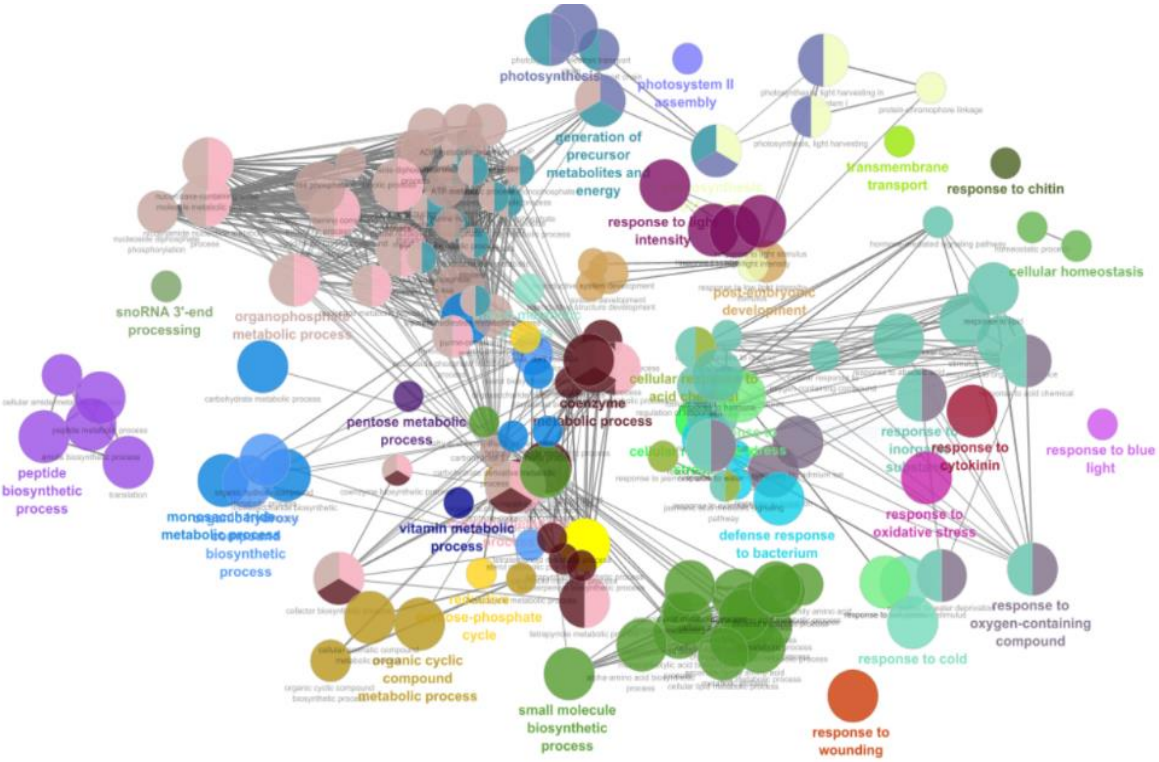
A)


%Genes / Term
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85
snoRNA 3'-end processing 5
cellular response to stress 123
Metabolism of carbohydrates 42
starch metabolic process 18
vitamin metabolic process 21
response to oxidative stress 78
response to wounding 54
response to blue light 20
response to cytokinin 49
response to chitin 27
photosystem II assembly 9
pentose metabolic process 8
transmembrane transport 91
response to temperature stimulus 96
response to cold 73
cellular homeostasis 49
homeostatic process 75
response to osmotic stress 106
response to salt stress 95
photosynthesis, dark reaction 8
reductive pentose-phosphate cycle 8
cellular response to acid chemical 78
response to jasmonic acid 41
jasmonic acid mediated signaling pathway 21
post-embryonic development 208
system development 237
reproductive structure development 164
reproductive system development 164
response to radiation 137
response to light stimulus 134
response to light intensity 49
response to high light intensity 34
response to external biotic stimulus 155
response to other organism 155
response to bacterium 80
defense response to bacterium 71
cellular aromatic compound metabolic process 490
heterocycle metabolic process 463
organic cyclic compound metabolic process 514
organic substance biosynthetic process 602
organic cyclic compound biosynthetic process 357
Metabolism 179
Arachidonic acid metabolism 15
Synthesis of Prostaglandins (PG) and Thromboxanes (TX) 9
Metabolism of lipids 58
Metabolism of steroids 22
photosynthesis, light reaction 45
photosynthesis, light harvesting 15

B)


% genes per group
organophosphate metabolic process 7.73% **
small molecule biosynthetic process 6.58% **
response to inorganic substance 9.22% **
generation of precursor metabolites and energy 2.73% **
coenzyme metabolic process 2.34% **
cofactor metabolic process 3.26% **
monosaccharide metabolic process 3.11% **
photosynthesis 2.45% **
response to oxygen-containing compound 6.71% **
organic hydroxy compound biosynthetic process 2.13% **
peptide biosynthetic process 1.94% **
photosynthesis, light reaction 1.04% **
snoRNA3'-end processing 0.11% **
cellular response to stress 2.62% **
photosystem II assembly 0.19% **
pentose metabolic process 0.17% **
transmembrane transport 1.94% **
Metabolism of carbohydrates 0.89% **
starch metabolic process 0.38% **
vitamin metabolic process 0.45% **
response to oxidative stress 1.66% **
response to wounding 1.15% **
response to blue light 0.43% **
response to cytokinin 1.04% **
response to chitin 0.58% **
response to cold 2.05% **
cellular homeostasis 1.6% **
response to osmotic stress 2.26% **
reductive pentose-phosphate cycle 0.17% **
cellular response to acid chemical 2.07% **
post-embryonic development 5.58% **
response to light intensity 2.92% **
defense response to bacterium 3.3% **
organic cyclic compound metabolic process 15.38% **
Metabolism 3.81% **

C)


photosynthesis
photosystem II assembly
generation of precursor metabolites and energy
transmembrane transport
response to chitin
response to light intensity
cellular homeostasis
snoRNA 3'-end processing
organophosphate metabolic process
post-embryonic development
coenzyme metabolic process
pentose metabolic process
response to cytokinin
response to blue light
peptide biosynthetic process
vitamin metabolic process
defense response to bacterium
response to oxidative stress
response to oxygen-containing compound
response to cold
organic cyclic compound metabolic process
small molecule biosynthetic process
response to wounding

D)

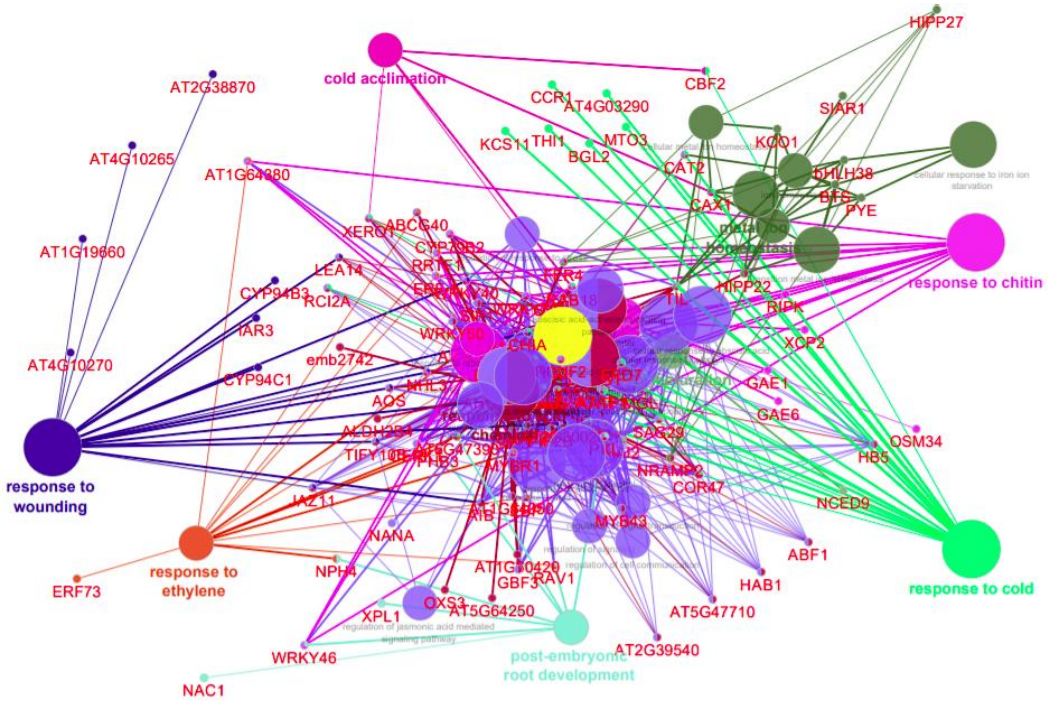
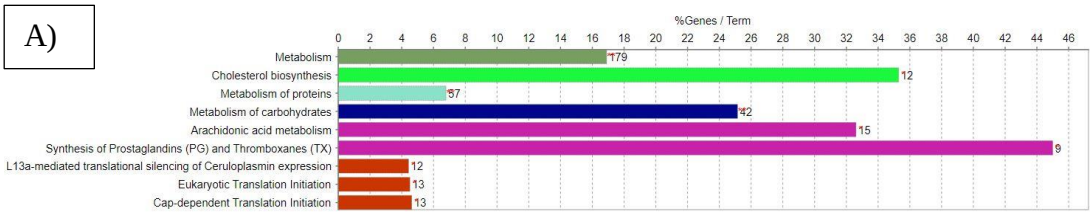


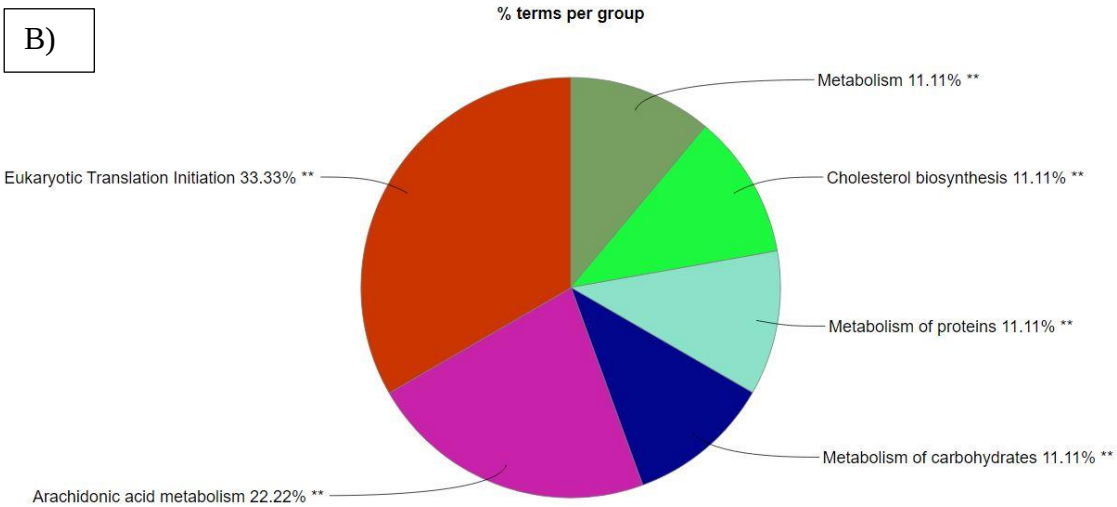
Figure 5.4: Network visualization of enriched terms among the differentially regulated genes during single drought stress. The network analysis was performed by ClueGo analysis. **a)** GO terms specific for solely drought DEGs from switchgrass. The bars represent the number of genes assigned with the terms. The percentage of genes per term is shown as bar label. **b)** Overview chart with functional groups including specific terms for DT DEGs. **c)** Over-represented GO analysis in the DT differentially expressed genes. These are functionally grouped network with terms as nodes linked based on their kappa score level (≥0.3), significant terms are shown. The node size represents the term enrichment significance. Functionally related groups partially overlap. The edges are related to the relationships between the selected terms defined based on the genes shared in a similar way. The label of the most significant term is used as the leading group term. **d)** Gene networks for GO biological process terms of DT DEGs.

5.4.5 Pathway analysis (Reactome) of drought (DT) differentially expressed genes

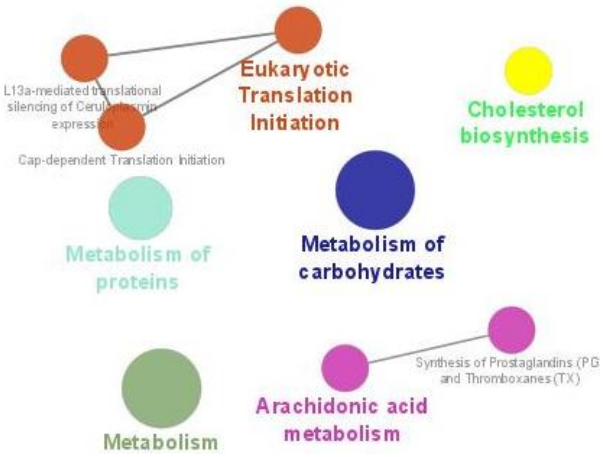
A)



B)



C)



D)

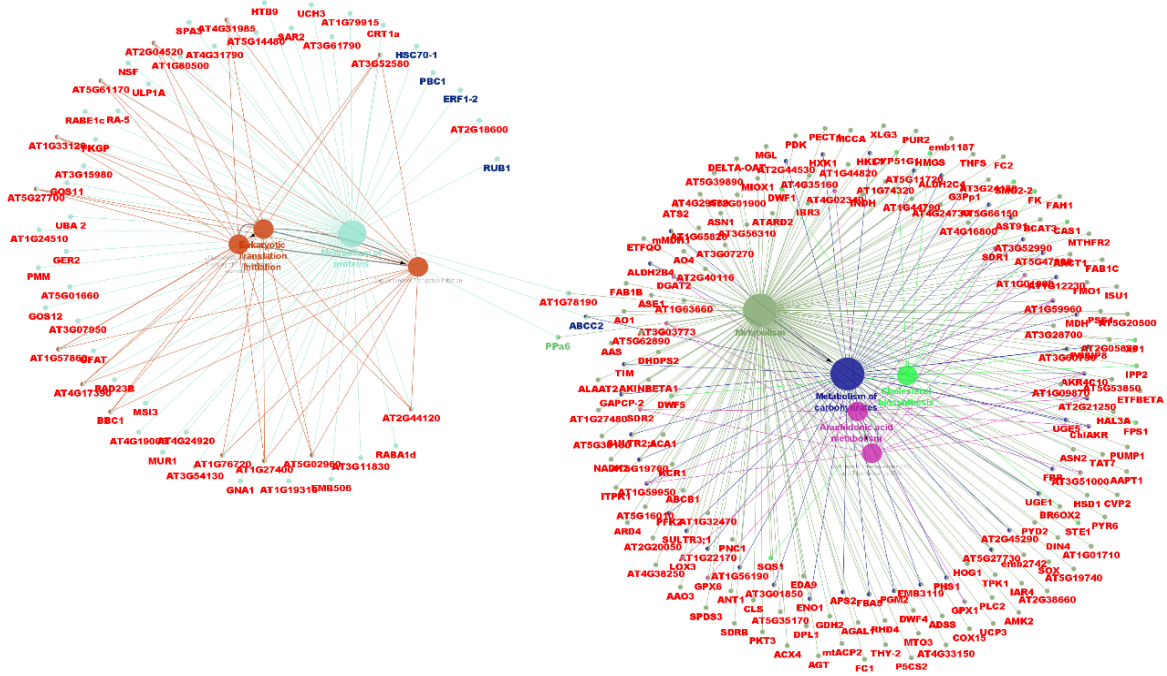
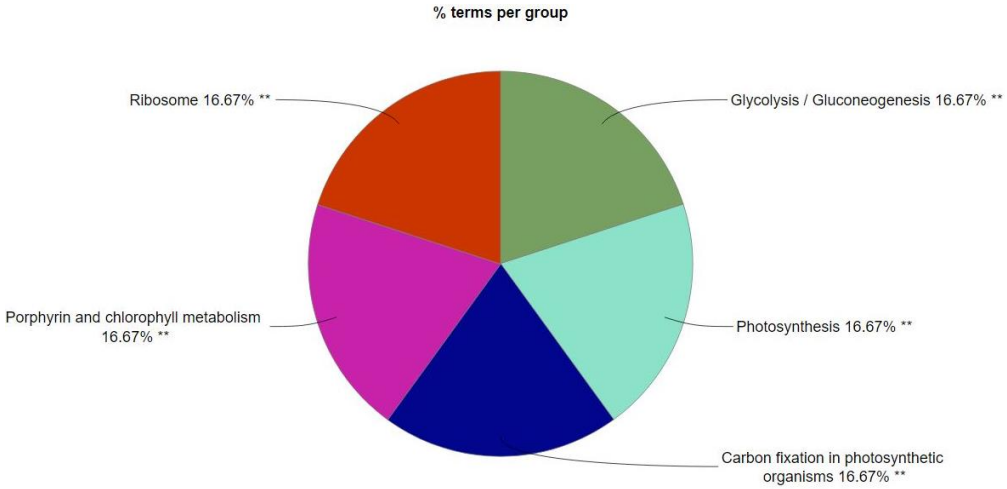


Figure 5.5: Network visualization of enriched pathways (Reactome) among the genes that were differentially regulated during single drought stress. The network analysis was performed by ClueGo analysis. **a)** pathway terms specific for only drought DEGs. The bars represent the number of genes assigned with the terms. The percentage of genes per term is shown as bar label. **b)** Overview chart with functional groups including specific pathways for DT DEGs. **c)** . Over-represented pathway analysis in the DT differentially expressed genes. These are functionally grouped network with terms as nodes linked based on their kappa score level (≥0.3), significant pathways are shown. The node size represents the term enrichment significance. Functionally related pathways partially overlap. The edges are related to the relationships between the selected terms which are defined based on the genes that are shared in a similar way. The label of the most significant term or pathway is used as the leading group term. **d)** Retrieved connection of the common genes of the pathway enriched by DT genes which include "metabolism and synthesis of prostaglandins (PG) and thromboxane (TX) is

5.4.6 Pathway analysis (KEGG) of DT DEGs



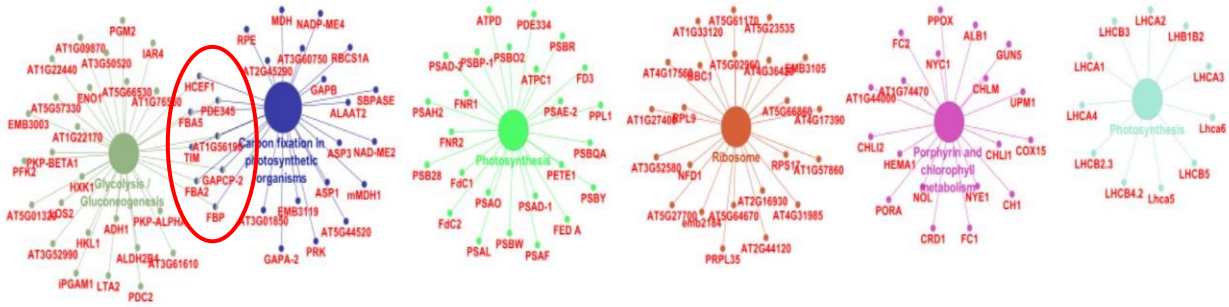


Figure 5.6: Pathway analysis (KEGG) of DT differentially expressed genes. The bars represent the number of downregulated DT genes assigned with the terms. The percentage of genes per term is shown as bar label. **b)** Overview chart of significant pathways for DT DEGs. **c)** Retrieved connection of the common genes of the significant pathway enriched by DT differentially regulated genes. Unique and shared genes between the pathways "glycolysis/gluconeogenesis" and "carbon fixation in photosynthetic organisms" have been circled in red.

The pathway enrichment analysis using Reactome database integration in ClueGO plug-in of Cytoscape software indicated "metabolism" as a common significant pathway enriched in the drought DEGs and combined drought and heat stress genes network analysis in **Figure 5.2c** and **Figure 5.5c**. However, the KEGG integrated pathway analysis in ClueGO indicated "carbon fixation and photosynthetic organisms" as the common enriched pathway for drought and combined drought and heat DEGs (**Figure 5.3c and Figure 5.6c**). The genes associated with each KEGG pathway identified are shown in the figures. These genes would be helpful for further analysis of stress tolerance study in plants. Carbon fixation and photosynthetic organisms and Glycolysis/gluconeogenesis pathways were identified from our KEGG analysis. These two major pathways have been previously reported in a drought-related response in sorghum [8]. The overlapping genes of the enriched terms carbon fixation and photosynthetic organism and glycolysis/gluconeogenesis as highlighted in red in Figure 5.3C are FBA2, FBP, PCK1, TP1, PGR1, PCK and HCEF1. These genes are potential candidate genes to be considered for drought and heat tolerance in plants. Interestingly, evidence of potential role of the gene "FBA2" in gluconeogenesis was confirmed from a literature collected in our MongoDB collection with PMID:22561114. Additional annotation of FBA2 from our collection indicate

that FBA2 could be located in the cytosol. Hence, these pathways could further be explored to understand stress mechanisms in plants.

5.5 Conclusions

The network analysis using Cytoscape provides additional enriched GO terms and pathways of stress-responsive genes in Chapter 2. In particular it provided the means to visualize the functional biological groups. Visualizing the network, with nodes representing GO terms or pathways and edges representing the relationship between these terms revealed the network's connectivity. Although the connections are computationally predicted, there is evidence to support the role of many of the genes that were highlighted to be enriched in the GO term or pathway. For example, ZIFL2 with shared terms in ‘ion transport’ and ‘stomatal closure’ contributes to drought tolerance by regulating stomatal closure. Possible novel relationships among pathways and terms could further be investigated using the unique or overlapping genes of the enriched terms or pathways.

REFERENCES

[1] Xia J, Gill EE, Hancock REW. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc 2015;10:823–44. https://doi.org/10.1038/nprot.2015.052.

[2] Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 2009;25:1091–3. https://doi.org/10.1093/bioinformatics/btp101.

[3] Paul Shannon 1, Andrew Markiel 1, Owen Ozier, 2 Nitin S. Baliga, 1 Jonathan T. Wang, 2 Daniel Ramage 2, Nada Amin 2, Benno Schwikowski, 1, 5 and Trey Ideker2, 3, 4 5, et al. Cytoscape: A Software Environment for Integrated Models. Genome Res 1971;13:426. https://doi.org/10.1101/gr.1239303.metabolite.

[4] Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: Which test? Bioinformatics 2007;23:401–7. https://doi.org/10.1093/bioinformatics/btl633.

[5] Rasmussen S, Barah P, Suarez-Rodriguez MC, Bressendorff S, Friis P, Costantino P, et al. Transcriptome Responses to Combinations of Stresses in Arabidopsis. Plant Physiol 2013;161:1783–94. https://doi.org/10.1104/pp.112.210773.

[6] Ge Y, Dudoit S, Speed TP, Glonek G, Solomon P, Grant GR, et al. Resampling-based multiple testing for microarray data analysis. Test 2003;12:1–77. https://doi.org/10.1007/BF02595811.

[7] Remy E, Cabrito TR, Baster P, Batista RA, Teixeira MC, Friml J, et al. A Major Facilitator Superfamily transporter plays a dual role in polar auxin transport and drought stress tolerance in Arabidopsis. Plant Cell 2013;25:901–26. https://doi.org/10.1105/tpc.113.110353.

[8] Woldesemayat AA, Ntwasa M. Pathways and Network Based Analysis of Candidate Genes to Reveal Cross-Talk and Specificity in the Sorghum (Sorghum bicolor (L.) Moench) Responses to Drought and It's Co-occurring Stresses. Front Genet 2018;9:1–22. https://doi.org/10.3389/fgene.2018.00557.

Chapter 6

SUMMARY

In summary, we generated gene expression data -> conducted data analysis -> provided annotation by homology -> and validated a candidate stress-responsive gene.

In this dissertation, first, we used transcriptomic data to investigate gene expression in switchgrass under a single drought and combinations of drought and heat stress (Chapter 2). This study contains lots of analysis, sequencing data, bioinformatics analysis, and rich information which provides new insights into abiotic stress response in switchgrass. Previous reports have shown that the regulatory mechanisms governing combined drought and heat stress are complex. The study's main findings are crucial for the full understanding of the study regarding abiotic stress response and tolerance. The different time points used in this study revealed genes and multiple phases of a stress response. Although heat stress alone was not imposed in the experiment, using bioinformatics analysis, heat-responsive genes were deduced using data from a single drought and combined drought and heat stress. This study's data on heat-responsive genes provide valuable information for validation in future heat tolerance studies in switchgrass and plants. Systems Biology involves co-expression analysis to identify a regulatory hub of genes. Using WGCNA, we grouped genes into biologically related groups as modules. A network visualization analysis using Cytoscape (Chapter 5) identifies enriched terms and pathways to support the GO enrichment and pathway analysis performed in Chapter 2. The overlapping genes from the network analysis serve as candidate genes to be considered for drought and heat

tolerance in switchgrass and plants in general. With the advent of high-throughput technology and increased availability of multi-omics data, an integrated approach combining different omics data is essential for a better interpretation. Combining omics data sets can help to understand the underlying mechanisms of the plant stress response. In a collaboration effort with the MGE lab at DSU, the same switchgrass tissues used for the RNA-Seq analysis in this study were used for epigenomic analysis by performing chromatin immunoprecipitation and sequencing (ChIP-Seq) analysis. We further studied how the DT- and DTHT-responsive peaks of the ChIP-Seq analysis correlate with the corresponding candidate genes identified in the RNA-Seq analyses. It is interesting to report that 155 DT responsive peaks overlapped with 118 DT responsive genes. Similarly, 121 DTHT responsive peaks overlapped with 110 DTHT responsive genes. The overlap of the epigenomic peaks and genes could be seen as plants' master regulators of the DT and HT genes. In the future, the biological function of genes identified in response to combined DT and HT stress could be confirmed by techniques such as single point mutation or RNAi. Genetic transformation using Agrobacterium strains has previously been reported in a lowland ecotype of switchgrass such as Alamo (same ecotype used in this study). Successful transformation enables gene function analysis and germplasm enhancement via gene editing biotechnology. This means that the putative PavirPAL1 which, has been cloned and functionally characterized, is ready for genetic transformation to improve switchgrass tolerance to stress. The stress-responsive genes, transcription factors, enriched GO terms, and pathways could be a basis for enhancimg biomass and bioenergy production of switchgrass.

Additionally, we established a pipeline (Chapter 4) to automatically retrieve scientific literature to study stress response in plants. The pipeline integrates databases (UniProt), text mining methods (PgenN, Textpresso), and literature resources (EuroPMC) to extract plant stress genes and link them to their function. We used the co-occurrence method of relation extraction of two concepts within the document, to indicate linkage. The data collected using the pipeline is used to find other processes of the differentially expressed genes in switchgrass (Chapter 2). The pipeline helped to extract new knowledge on plant stress response to complement existing knowledge in databases and other plant resources. Additional evidence to support the role of the stress-responsive genes from the gene expression analysis was obtained using the pipeline. The pipeline developed serves as a framework to retrieve scientific literature to study other organisms not just plants by integrating relevant resources.

This work generated multiple testable hypotheses. For example, a candidate gene, Phenylalanine ammonia-lyase 1, the first enzyme in the phenylpropanoid pathway that catalyzes the deamination of $_{L}$-phenylalanine to trans-cinnamic acid, was validated in switchgrass. PAL-genes are predominantly found in plants and have been identified and characterized in a number of plants species but underexplored in switchgrass. For example, Arabidopsis has four PAL-genes (: **AT2G37040 (PAL1)**, AT3G53260, (PAL2), AT5G04230 (PAL3) and AT3G10340 (PAL4). A section of this study provides the first report of characterizing a PAL gene in switchgrass. Switchgrass has eleven PAL genes. Four out of the eleven genes were downregulated in the RNA-Seq of the switchgrass transcriptome under DT and HT stress. This study serves as a platform to further understand the function of each of the PAL-genes in

switchgrass. The co-occurrence method achieves a relatively high recall. A shortcoming of this extraction method is that the approach can ignore the context of each co-occurrence, leading to low precision. In the future, a complex relation extraction method such as Natural Language Processing can be used to retrieve information on the relation between stress genes and their function. Furthermore, we can integrate the pipeline described in the iTextMine framework in the future. iTextMine integrates text mining tools to extract comprehensive knowledge from the scientific literature.

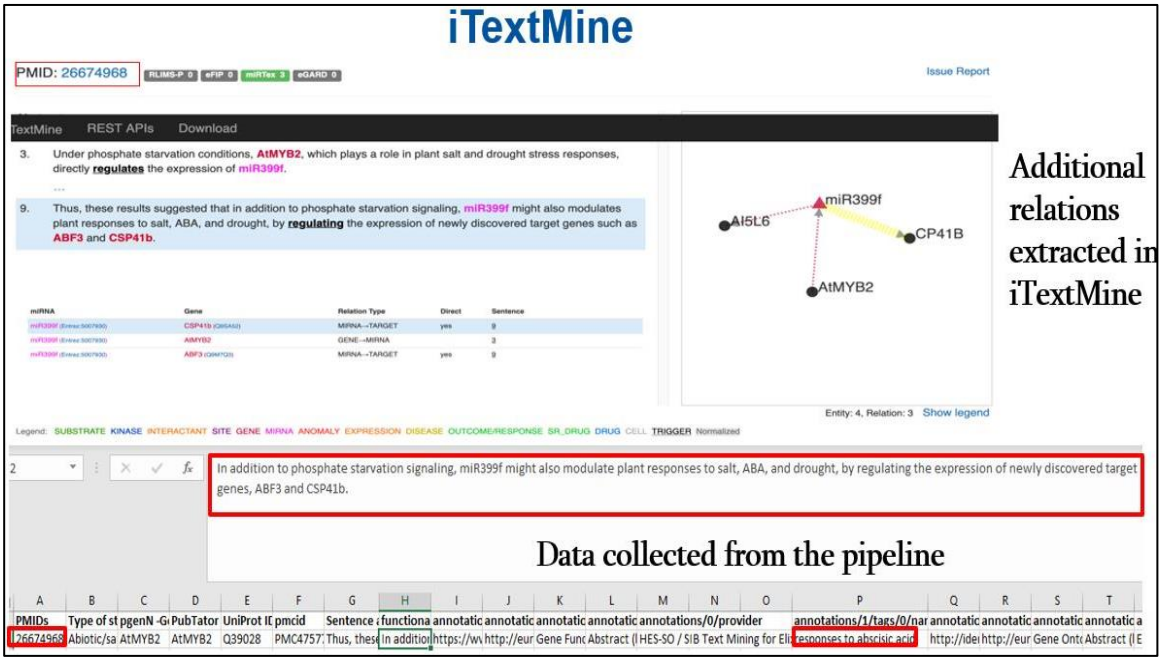


Figure 6.1: Captured in iTextMine MYB2 which is one of the stress genes collected. The figure shows MYB2 regulation of microRNA involved in abscisic acid response. Through an experiment MYB2 is upregulated in stress and by using iTextMine we could identify the underlying mechanisms.

Data curation: An additional outcome or application of the pipeline is a set of publications for UniProt entries with annotations that can be submitted to the UniProt Knowledgebase via the community submission system. UniProt is a publicly available database offering sequence and functional annotation for proteins across all taxonomic groups (https:///www.uniprot.org). UniProtKB consists of two sections: the reviewed (Swiss-Prot) section consisting of expert-curated entries; and the unreviewed (Tremblor) section that provides automatic annotations and represents approximately 99 % of the entries. Expert curation is time-consuming, and UniProt has annotation priorities and a limited biocuration task force. Thus, additional mechanisms to increase the entry information content are key to keep the resource up-to-date. The community submission system allows UniProt users to add publications and annotations to proteins of their interest. An example of submitting a publication and annotation of a protein of interest; Phenylalanine ammonia lyase 1 is shown below.

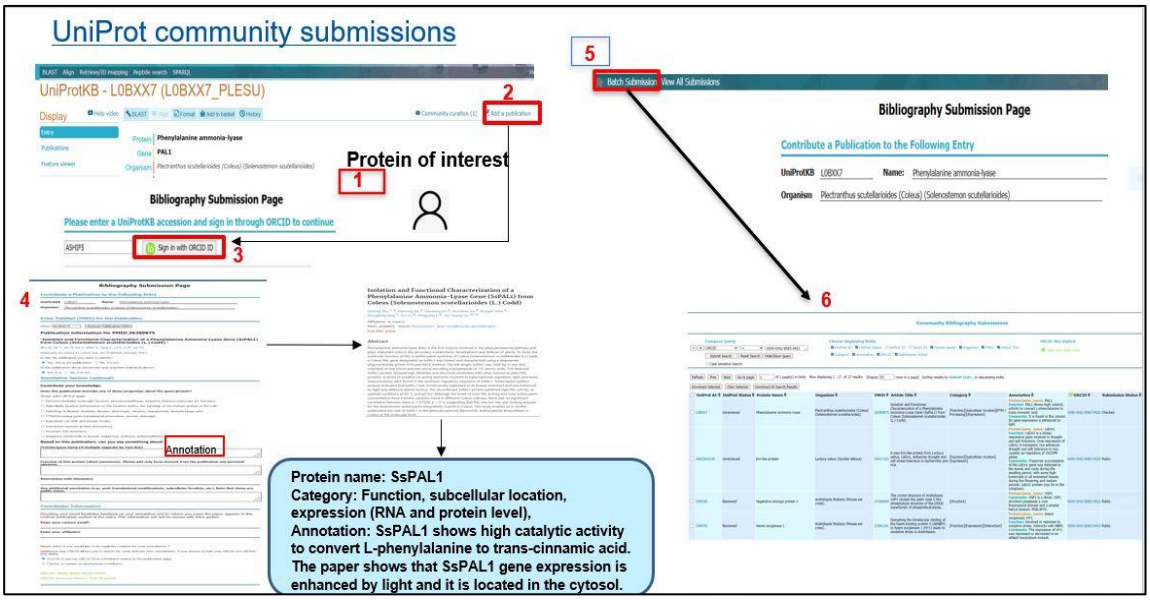


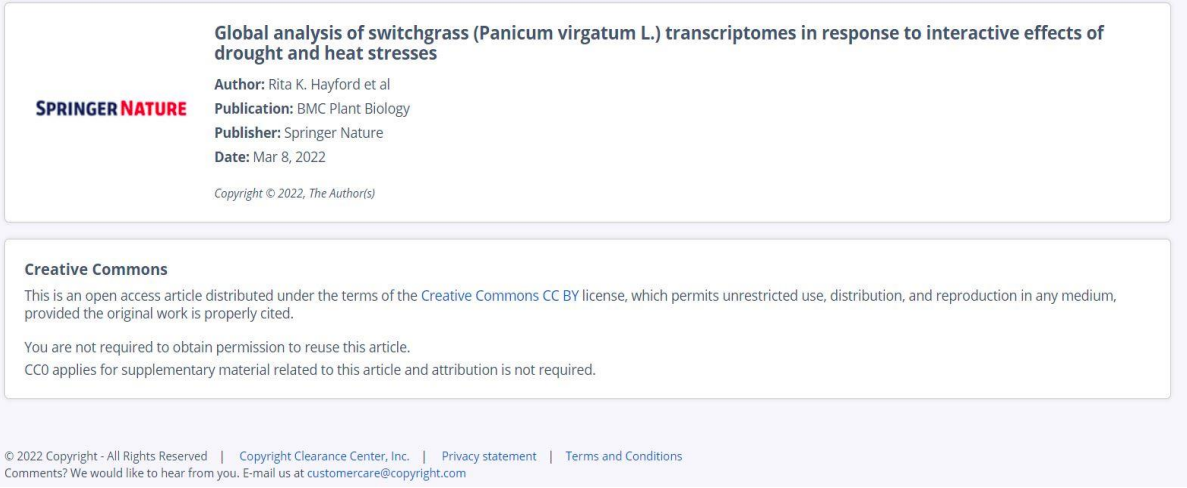
Figure 6.2: Submission of annotation of PAL1 to the UniProt database.

PAL1 was differentially expressed under stress and the pipeline allowed to bring data from other species from literature that supports its involvement in stress. PAL1 from *Solenostemon scutellarioides* (SsPAL1) gene expression is enhanced by light and located in the cytosol. We then pursued the functional validation of PAL1 in switchgrass. Additional annotation and publication of over 20 proteins involved in stress have been added to the UniProt database. The MongoDB collection can be reviewed further to add new annotation of stress genes and publications that are not already in UniProt.

Appendix

PERMISSIONS

Chapter 2 is published in 2022 BMC Plant Biology and is an open access articles distributed under the terms of the Creative Commons Attribution License 4.0 International License which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons License, and indicate if changes were made. The images or other third-party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly form the copyright holder.



SPRINGER NATURE

Global analysis of switchgrass (Panicum virgatum L.) transcriptomes in response to interactive effects of drought and heat stresses

Author: Rita K. Hayford et al
Publication: BMC Plant Biology
Publisher: Springer Nature
Date: Mar 8, 2022

Copyright © 2022, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.
CC0 applies for supplementary material related to this article and attribution is not required.