FUNCTIONAL AND EVOLUTIONARY DYNAMICS OF GENES INVOLVED IN DROUGHT TOLERANCE IN LOBLOLLY PINE (*PINUS TAEDA* L.)

A Dissertation

by

JINGJIA LI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Claudio Casola
Committee Members,	Carol Loopstra
	Hongbin Zhang
	James Cai
Head of Department,	Kirk Winemiller

December 2020

Major Subject: Ecosystem Science and Management

ABSTRACT

Drought, a major threat to the health and productivity of both natural ecosystems and agriculture, is expected to increase in frequency and intensity across many regions as a consequence of climate change and repurposing of natural water resources. Loblolly pine (Pinus taeda L.) represents a major forest species across the southeastern US due to its widespread distribution, ecological prominence, and extensive utilization for the industrial production. Thus, developing loblolly varieties with increased tolerance to aridity is a major goal of the forest industry. However, this will require a significant leap forward in our understanding of the genetic basis of drought tolerance in loblolly. The main goal of this project is to generate genomic resources and bioinformatic approaches to identify genes, regulatory regions and genetic variants involved in drought tolerance in loblolly pine. In the first component, I analyzed transcriptomic (RNA-seq) data from two loblolly genotypes with divergent tolerance to aridity. I identified more than 4,000 drought-related transcripts in response to drought in the root of Pinus taeda. Genotype x Environment (GxE) interactions were prevalent, suggesting that very different cohorts of genes are influenced by drought in the tolerant vs. sensitive loblolly genotypes. In the second part, I identified nearly 9,500 unique sites representing 24 clusters of Transcription Factor Binding Sites (TFBSs) in the promoter region of 1,386 DRTs. All of the 24 TFBSs share homology with known motifs in flowering plants. A total of 1,046 unique DRTs linked to 16 TFBSs were associated to 213 overrepresented non-redundant GO terms, most of which are related to processes known to be involved in drought

tolerance. In the third component of my research, I integrated the transcriptome data with extensive genetic variant (SNP) datasets in loblolly to determine the evolutionary dynamics associated with DRTs. I found that DRTs share higher rates of adaptive evolution and contain a higher than expected number of SNPs associated with aridity than other genes. Overall, these findings will assist the sustained effort to develop varieties of loblolly pine that can better sustain the projected increase in aridity along the range of this key forest species.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Claudio Casola, for his constant encouragement and prominent guidance on the way of pursuing my doctoral Degree. I am deeply indebted to his patient advising, friendly help, and continuous support to my research. I would also like to extend my deepest gratitude to my committee members, Dr. Carol Loopstra, Dr. Hongbin Zhang, Dr. James Cai, for their guidance and support throughout the course of this research.

I am grateful to the McMillan Ward Memorial Foundation for supporting my graduate work.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

I would like to appreciate the previous and current members in Dr. Casola's lab, Tomasz Koralewski, Weixi Zhu, Shelby Landa, Xuan Lin for their generous assistance on my research and life.

Finally, I take this opportunity to than my mother and father for their encouragement and to my husband for his patience and love. Without their enduring love and support, I could not have done this. I dedicate the dissertation to them.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Claudio Casola (Department of Ecology and Conservation Biology), Dr. Carol Loopstra (Department of Ecology and Conservation Biology), Dr. Hongbin Zhang (Department of Soil and Crop Sciences) and Dr. James Cai (Department of Veterinary Integrative Biosciences).

The physiological measurements used in Chapter 2 were conducted by Dr. Jason West (Department of Ecology and Conservation Biology), and the RNA-Sequencing was prepared by Jeff Puryear and sequenced by the Genomics & Bioinformatics Service at Texas A&M University. The annotation data with EnTAP for the loblolly transcriptome was provided by Dr. Jill Wegrzyn (Department of Ecology and Evolutionary Biology, University of Connecticut). Part of the data analyzed for Chapter 4 was provided by Dr. Carol Loopstra and Dr. Mengmeng Lu.

Funding Sources

Graduate study was supported by the McMillan Ward Memorial Graduate Fellowship and the Department of Ecosystem and Science Management. Clones of the loblolly pine varieties used in this study were provided by ArborGen Inc. This work was supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number TEX0-1-9599, the Texas A&M

AgriLife Research, the Texas A&M Forest Service and the Western Gulf Forest Tree Improvement Program Cooperative. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

Drought is a severe problem across multiple ecosystems and it is expected to increase in frequency and intensity in some areas due to climate change and altered watershed use. Drought features can be influenced by multiple aspects, for instance, circulation patterns, evapotranspiration, and air temperatures; regardless, drought represents a natural hazard to many ecosystems (BUCHANAN - SMITH AND WILHITE 2005; (IPCC) 2013). Because of the generation time of most tree species, forests are likely to be critically affected by drought. In the United States alone, forest ecosystems occupy about one-third of land surface and store nearly half of the carbon found in terrestrial ecosystems (BONAN 2008; AGRICULTURE 2016). The southeastern states, including Texas, harbor a significant proportion of forestland in the continental US, which is poised to become increasingly arid in the next few decades. For example, climate projections for the years 2021-2065 show that in east Texas mean annual temperatures, warm and dry spells and number of days/year with minimum temperature above 20°C will increase, whereas precipitation will decrease in this region. These changes in climate regime are likely to induce a significant loss of productivity and tree mortality (BRESHEARS *et al.* 2005; VAN MANTGEM AND STEPHENSON 2007). Loblolly pine (*Pinus taeda* L.) represents the most important species for the forest industry both in Texas and across the southeastern US. This important conifer is native to North America from New Jersey to Florida and Texas. Loblolly pine forests occupies 55 million acres, or about one-fourth of the southern forests in the U.S, (W. BRAD SMITH

2007). Along the loblolly range, annual precipitation historically has ranged from 40 to 50 inches/year (1,020-1,270 mm) (WAHLENBERG 1960). Natural loblolly forests contribute important ecosystem services, from carbon sequestration (JOHNSEN 2004) to the support of wildlife, including the endangered red-cockaded woodpeckers and a variety of other birds and mammals (WAHLENBERG 1960). Loblolly also represents one of the most important commercial forest crops in North America due to its rapid growth and high productivity, contributing to nearly 80% of all cultivated trees and about half of the wood products generated by forest products industry in the southeastern US (W. BRAD SMITH 2007; GREIS 2013).

It has been demonstrated that low water availability due to drought affects multiple aspects of loblolly biology. As typical of most plants, low soil moisture is associated with reduced (SCHMIDTLING 2001) to arrested growth (GRISSOM 1997) and increased mortality rates, as observed in the 2011 exceptional drought season in Texas (KLOCKOW *et al.* 2020). Specific phenotypic traits, including important commercial traits, can also be impacted by plant dehydration, such as branch growth, needle length, and ring width (GRAHAM *et al.* 2012), and sensitivity to pathogenic fungi such as *Leptographium terebrantis* (PRATIMA DEVKOTA 2018). Locally adapted loblolly genotypes with varying levels of sensitivity to aridity have been described, with varieties native of regions with higher precipitation typically showing higher productivity but lower tolerance to low moisture (PRISLEY 2019). Given the impact of drought on loblolly forest health and productivity and the ongoing changes in climate regime, efforts to understand the genetic mechanisms implicated in drought tolerance in loblolly pine have

become increasingly important. These endeavors have the potential to improve strategies in loblolly pine management and breeding by identifying both genetic markers associated with resistance to lower moisture regimes and genes involved in the processes that are more severely compromised by aridity.

Plant response to drought occurs across numerous traits at several organizational scales (e.g., cellular, tissue, whole plant). These responses are linked to a wide range of genes that are differentially expressed across these scales. The genotypic component depends on natural selection and adaptation, fundamental physiological or morphological tradeoffs, and other drivers that affect gene expression and environmental responses (CHAVES *et al.* 2009). Understanding how species like loblolly pine responds to drought therefore requires interdisciplinary efforts that integrate these components. These efforts should also enhance our ability to identify improved selection strategies and aid in forecasting forest responses to climate change.

The genetic basis of drought tolerance in loblolly has been analyzed using different approaches. For instance, a genetic component to the responsiveness of xylem morphology and leaf-level physiology to drought has been identified (SPERRY *et al.* 2002). Studies based on large-scale transcriptomic and genetic data have revealed some components of the genetic networks involved in drought response of this pine species (LORENZ *et al.* 2011) and other conifers (MORAN *et al.* 2017). However, the genetic bases of drought tolerance in loblolly pine are still largely unknown. For example, it is not clear which genes are involved in drought tolerance, which regulatory regions are shared between drought-related genes in conifers, and if these regions are conserved

with respect to angiosperm genes involved in the response to drought. Furthermore, these studies suffer from the evolutionary distance between conifers and angiosperms. Functional gene annotation in conifers is still very limited and many drought-related genes have no apparent functional equivalent in angiosperms (PRUNIER *et al.* 2016). A primary goal of my dissertation research is to contribute to the general understanding of the genetic basis of response and tolerance to low water availability in loblolly pine. In this study, I have integrated novel transcriptomic datasets, *de novo* discovery of transcription factor binding sites (TFBSs), and selection regime on drought-related genes to achieve three objectives:

1) Identifying genes and genetic networks involved in drought tolerance among loblolly pine varieties.

2) Identifying regulatory motifs associated with drought-related genes that are upregulated and downregulated in response to drought.

3) Assessing signatures of adaptation that shaped the evolution of drought-related genes in ~370 loblolly pines sampled across the range of this species.

The genetic response to drought is primarily associated to changes in the expression of a large suite of genes. Interspecific variation in this response is common and associated with drought tolerant and sensitive genotypes. The extent to which different genetic networks orchestrate the adjustments to water deficit in tolerant and sensitive genotypes has not been fully elucidated, particularly in nonmodel plants. In loblolly pine, studies on gene expression changes induced by drought stress have been conducted in the last two decades using either microarray-based techniques (HEATH *et*

al. 2002; WATKINSON *et al.* 2003; LORENZ *et al.* 2011; MICHAEL *et al.* 2020) or expressed-sequence tags (ESTs) data (LORENZ *et al.* 2006).

In the first part of my dissertation (Chapter 2), I performed RNA-sequencing analyses of root tissues exposed to simulated drought conditions from two clones with contrasting tolerance to drought and assembled de novo transcriptome from the RNA-sequencing of loblolly ramets. I found significant changes in expression levels in more than 3,500 drought-related genes. Because most differential expression and subsequent analyses involved transcripts rather than individual genes, the focal genetic units of my project are represented by drought-related transcripts (DRTs) rather than drought-related genes. I found that Genotype x Environment (GxE) interactions were prevalent, suggesting that very different cohorts of genes are influenced by drought conditions in the tolerant vs. sensitive genotypes.

In the second component of my project (Chapter 3), I investigated the composition of DRT promoter regions and identified nearly 9,500 sites representing 24 clusters of unique TFBSs. These short *cis*-regulatory motifs dictate the timing and duration of transcription through their interaction with transcription factors. This represented the first large-scale computational analyses of TFBSs among gymnosperms. A major finding of this analysis is that all of the 24 TFBSs found in loblolly DRTs are homologous with known motifs described in flowering plants.

In the third section of my dissertation (Chapter 4), I tested the hypothesis that DRTs experience more rapid adaptive evolution than other genes. Previous studies based on population genomic datasets in loblolly have shown that several genetic variants

(SNPs) are associated with either aridity of environmental variables that are related with low water variability (ECKERT *et al.* 2010; DE LA TORRE *et al.* 2019; LU *et al.* 2019). I found that these variants occur in DRTs at a significantly higher frequency than expected based on other genes. Using more than 2.8 million SNPs identified by exome-capture and sequencing in Dr. Carol Loopstra's laboratory (LU *et al.* 2016; LU *et al.* 2017), I also found that, overall, DRTs experience higher rates of adaptive evolution than other genes.

The results of my dissertation project provide the most comprehensive analyses of drought-related genes in *Pinus taeda*, and one of the most extensive works on the genetic basis of drought tolerance in gymnosperms. Through the integration of transcriptomic, *cis*-regulatory and adaptation datasets, I have shown that remarkably different genetic networks are involved in the response to drought between loblolly varieties. The identification of an array of TFBSs conserved between loblolly and angiosperms implies that, surprisingly, many *cis*-regulatory motifs are shared between distantly related seed plants. Finally, I validated the hypothesis that drought-related genes are evolving rapidly. These findings will enable a better understanding of loblolly pine adaptation to drought, improve the ability to develop aridity-tolerant loblolly varieties through breeding, and prompt further research on the evolution of regulatory regions and the action of natural selection on stress-related genes in seed plants.

2. EXTENSIVE VARIATION IN DROUGHT-INDUCED GENE EXPRESSION CHANGES BETWEEN LOBLOLLY PINE GENOTYPES

2.1. Introduction

Low water availability affects productivity and growth in natural forests and in tree plantations and is expected to become a primary limiting factor in certain areas due to local climate shifts (KARL *et al.* 2009). The combination of decreased precipitation and higher temperatures is predicted to exert a strong selective pressure on natural tree populations. Plant response to drought occurs across numerous traits at several organizational scales (e.g., cellular, tissue, whole plant). These responses are linked to a wide range of genes that are differentially expressed across these scales. The genotypic component depends on adaptation to local environmental conditions, fundamental physiological or morphological tradeoffs, and other factors that affect gene expression and environmental responses (CHAVES *et al.* 2009). These factors play a significant role in variation in drought tolerance within populations, particularly those of species with broad ranges, wherein genotypes with high and low tolerance to aridity can evolve in response to the local climate. Thus, investigating the genetic basis of drought tolerance in species with populations adapted to a variety of water availability conditions is an essential approach to determine how plants respond to this type of abiotic stressors. Species with large population sizes and locally adapted varieties might better sustain climate changes throughout the migration of drought-tolerant genotypes towards areas that will become increasingly more prone to water deficit (AITKEN *et al.* 2008).

Loblolly pine (*Pinus taeda* L.) represents the most commonly planted trees across the southeastern United States (HAMBERGER *et al.* 2009). Local adaptation in loblolly pine has been documented by a number of studies on several phenotypic traits (ECKERT *et al.* 2010; QUESADA *et al.* 2010; CUMBIE *et al.* 2011; PALLE *et al.* 2011), including tolerance to aridity (EVENO *et al.* 2008; ECKERT *et al.* 2010). For example, Eckert et al. identified 5 loci associated with levels of aridity in *P. taeda* using 3,059 SNPs (ECKERT *et al.* 2010). Large-scale datasets of polymorphisms have recently become available in loblolly via exome-based genotyping analyses, enabling the identification of a high number of polymorphisms associated with traits, climate variables or genes known to be involved in drought tolerance. Genotype-phenotype association studies based on these data have revealed a few SNPs and SNP-SNP epistatic interactions associated with $\Delta^{13}C$, a proxy of water use efficiency that might be related to drought tolerance (LU *et al.* 2017). Additionally, 611 unique SNPs were found to be associated with 56 climate and geographic variables, including several hundred SNPs associated with temperature and precipitation variables, some of which might correlate with drought tolerance (LU *et al.* 2019). The combined analysis of exome polymorphisms, gene expression and metabolomic data has also shown 661 SNPs associated with drought-related genes (LU *et al.* 2018). Using 87,000 SNPs obtained from genome resequencing data (DE LA TORRE *et al.* 2018), De La Torre and collaborators also reported that water availability represents the primary climate variable associated with local adaptation in loblolly (DE LA TORRE *et al.* 2019).

A complementary approach to identify genes associated with drought tolerance

consists in assessing variation in gene expression in controlled experiments, including water-deficit stress treatments of genotypes with varying tolerance to aridity. This approach has revealed that the expression level of thousands of genes from a multitude of genetic networks is significantly affected as a result to prolonged low water availability (OSAKABE *et al.* 2014). In loblolly, studies on gene expression changes induced by drought stress have been conducted in the last two decades using either microarray-based techniques (Heath et al. 2002; Watkinson et al. 2003; Lorenz et al. 2011) or expressed-sequence tags (ESTs) data (Lorentz et al. 2005). Overall, genes with similar functions have been found to be over- or underexpressed in both flowering plants and gymnosperms. These genes are involved in an array of cellular processes activated by drought stress, including protection from oxidative-, heat- and osmotic-stress, changes in metabolic functions, transcription regulation and release of hormones and other signaling molecules (MORAN *et al.* 2017). Similar results have been reported in microarray or transcriptomic studies of other drought-stressed conifers, including *Pinus pinaster* and *Pinus pinea* (PERDIGUERO *et al.* 2013), *Pinus halepensis* (FOX *et al.* 2018b), *Abies alba* (BEHRINGER *et al.* 2015), *Pseudotsuga menziesii* (MULLER *et al.* 2012) and *Cunninghamia lanceolata* (HU *et al.* 2015).

Early studies in loblolly pine seedlings exposed to drought have shown expression changes in genes encoding S-adenosylmethionine synthetase, transcription factors belonging to the ABA pathway, glycoproteins and glycine-rich protein associated to the cell wall (CHANG *et al.* 1996). Further works have pointed to changes in the activity of genes encoding stress-response proteins, including heat shock proteins,

dehydrins and other late embryogenic-abundant (LEA) proteins, as well as enzymes involved in several metabolic pathways (WATKINSON *et al.* 2003; LORENZ *et al.* 2006). In one of the most comprehensive analysis of gene expression in drought-stressed loblolly, Lorenz and co-authors identified multiple genetic networks involved in drought response, including 9-cis-epoxycarotenoid dioxygenase, zeatin O-glucosyltransferase, and ABA-responsive protein (LORENZ *et al.* 2011). Analogous investigations in other conifers have largely mirrored these findings (MORAN *et al.* 2017). Importantly, the expression level of these genes was comparable in control and drought seedlings following re-watering of water stressed plants (WATKINSON *et al.* 2003; LORENZ *et al.* 2006; LORENZ *et al.* 2011).

Variation in gene expression between loblolly genotypes in response to drought stress has also been described. For instance, LORENZ *et al.* isolated and analyzed the expression of 6,765 partial transcripts obtained from the root of three unrelated loblolly genotypes in control, drought stress and drought recovery regimes. In this study, 110 transcripts changed expression by genotype, compared to 42 transcripts with variation due to treatment. While these findings suggest that genetic variation plays a major role in the differential response to drought across loblolly populations, they were obtained from a limited subset of partial transcripts expressed in root tissues. To provide a comprehensive description of the genes involved in drought response in loblolly, we performed a transcriptomic analysis of control and drought-stressed root systems from two loblolly clones with different physiological responses to drought. Physiological traits such as growth, soluble carbohydrate, $\delta^{13}C$, water potential, gas exchange

measurements, specific leaf area and leaf nitrogen content have shown differences in the water relations between these two clones.

We found more than 4,000 transcripts with significant changes in expression level in seedlings grown under drought conditions in either clone. Few of these drought-related transcripts were shared between the clones, indicating extensive genotype by environment interactions between these drought tolerant and sensitive loblolly genotypes. Although GxE interactions were less prevalent at the level of functional gene annotations (GO terms) and metabolic pathways, they were common among transcription factors and transcription factor families encoded by drought-related transcripts. These findings revealed an unexpected divergence in the genetic networks involved in the response to water deficit between loblolly genotypes.

2.2. Results

2.2.1. Physiological Measurements of Drought Effects in Loblolly varieties

We analyzed ramets from three loblolly pine clones in randomized experimental greenhouse plots with two water treatments, herein referred to as control and drought. We found significant variation in traits including water potential between the two treatments as well as remarkable differences between clones under the same water regime (**Figure 2.1**). Clone 2 and clone 5 showed the most prominent difference water deficit tolerance and were selected for subsequent transcriptomic analyses. Further analyzed traits included hydraulic conductivity, P_{50}, wood density, $\delta^{13}C$, root biomass, leaf nitrogen and $\delta^{15}N$.

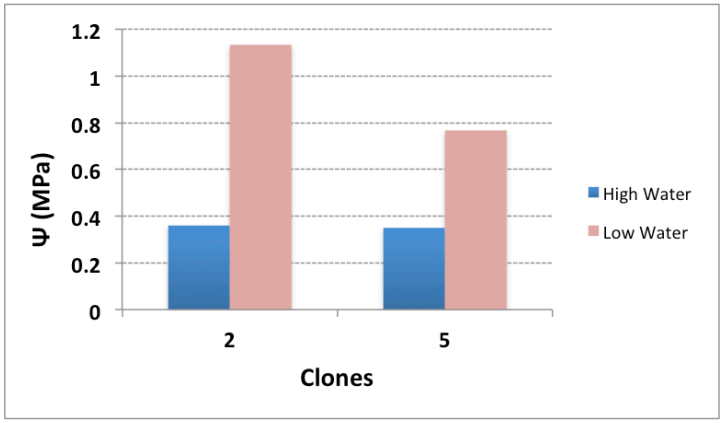


Figure 2.1 Water potential in control and drought-simulated ramets across the two loblolly clones 2 and 5.

2.2.2. Transcriptome Assembly from Loblolly Pine Root and Needle samples

Twenty-four total RNA samples were isolated, processed and sequenced as described in the Materials and Methods. A needle-library was discarded due to high bacterial contamination. A total of 99,756 transcripts were mapped to the loblolly pine v.1.01 genome and assembled from reads of the 23 remaining RNA-seq libraries using the HISAT2 and StringTie tools (KIM *et al.* 2015; PERTEA *et al.* 2016). Given the size and high redundancy of the loblolly genome we applied stringent mapping conditions to remove reads aligned to multiple loci and reads with more than 2 mismatches with the genome (Materials and Methods). TransDecoder was applied to detect candidate protein coding regions from the assembled transcripts (TANG *et al.* 2015). Approximately 60% of transcripts showed protein-coding capacity given the conditions set to identify open reading frames (**Materials and Methods**). A total of 54,826 transcripts were considered protein-coding according to TransDecoder, 53,256 of which were expressed in the root

and were used.in the following analyses.

2.2.3. Genetic distance between the clones and the reference genome

Clones with different genetic distances from the reference genome could lead to a bias in the transcript abundance quantification because of the different probability of mapping reads between clones. However, we found no significant difference in the genetic distance between the libraries of the two sequenced clones and the reference genome for root tissues (P-value: 0.55 for needle and 0.61 for root). Accordingly, the proportion of mapped reads was comparable between the two clones after removing an outlier library in clone 5 with much higher number of mapped reads. Moreover, we observed a similar number of transcripts between the two clones for the root tissues compared to needles.

2.2.4. Transcriptome Response to Simulated Drought in Loblolly Pine Root

Differentially expressed genes between drought and control conditions and between clones were identified using DESeq2 (LOVE *et al.* 2014) with applying the threshold value of log-fold change at 1 and the expression difference at 5% FDR. Genes that were differentially expressed between drought and control experiments were defined drought-related transcripts or DRTs. Using expression levels from root and needle libraries, we identified 4,012 and 29 DRTs in the two organs, respectively (**Table 2.1**). This corresponds to 7.9% and 0.07% of the total transcripts annotated in the root and the needle, respectively. The expression of 12 root DRTs and 10 needle DRTs were further

analyzed using qRT-PCR. There were 8 upregulated DRTs and 4 downregulated DRTs included in the root samples. There are 7 upregulated DRTs and 3 downregulated DRTs conducted in needle samples. We found a strong positive correlation between RNA-seq and qRT-PCR results between drought and control in root, whereas needle samples showed a much lower correlation. Given the low number of DRTs found in the needle and the limited correlation between RNA-seq and qRT-PCR data, we focused exclusively on the root data in the remainder of the study.

Similar numbers of upregulated and downregulated DRTs were observed in the root; however, clone 5 showed remarkably more upregulated DRTs compared to clone 2 (**Figure 2.2**; **Table 2.1**). Unexpectedly, the two clones also exhibited very little overlap of their DRTs: only 6-13% of upregulated and 10-11% of downregulated DRTs overlapped between clones 2 and 5 (**Figure 2.2**; **Table 2.1**). Furthermore, a higher number of clone-specific transcripts were found in clone 5, especially upregulated ones, compared to clone 2 (**Table 2.1**, "Only clone 2" "Only clone 5"). In total, we identified only 87 upregulated DRTs and 108 downregulated DRTs shared between clones. In addition, 17 DRTs showed opposite expression patterns between clones, 14 of which were upregulated in clone 5 and downregulated in clone 2 (**Figure 2.2**; **Table 2.1**, "Clones 2 and 5 opposite"). We also identified 802 clone-specific DRTs with opposite expression patterns between clones. The average difference in LFC ($\log_2$ fold change) between clones for these 819 transcripts was 6.1.

Table 2.1 Root up- and downregulated DRTs and non-DRTs

	DRTs		non-DRTs	
	Upregulated	**Downregulated**	**Upregulated**	**Downregulated**
Clone 2	662	1041	22,105	21,591
Clone 5	1391	981	23,038	21,563
Both clones combined (bcc)	362	507	23,262	22,802
Clones 2 and 5 opposite	3	14	7,332	7,469
Only clone 2	405	718	196	195
Only clone 5	1223	773	381	281
Only bcc	43	106	0	0
Only clones 2 and 5	2	5	0	0
Only clone 2 and bcc	167	201	4,960	5,405
Only clone 5 and bcc	67	97	4,673	4,287
All combined	85	103	13,473	12,859
Total	**2009**	**2020**	**31,803**	**30,522**

Clone 2: total DRTs in clone 2; Clone 5: total DRTs in clone 5; bcc: both clones combined; clone 2 and clone 5 opposite: up- or downregulated DRTs in clone 2 shown to be corresponding opposite regulation in clone 5; only clone 2: DRTs shown only in clone 2; only clone 5: DRTs shown only in clone 5; only bcc: after getting the DRTs form bcc dataset, the DRTs shown in only one clone; only clone 2 and 5: DRTs common in clone 2 and clone 5 but not overlapped with DRTs from dataset when combine the two clones; only clone 2 and bcc: DRTs in common between only clone 2 and bcc; only clone 5 and bcc: DRTs in common between only clone 5 and bcc; all combined: DRTs in common among clone 2, clone 5 and bcc in each regime.

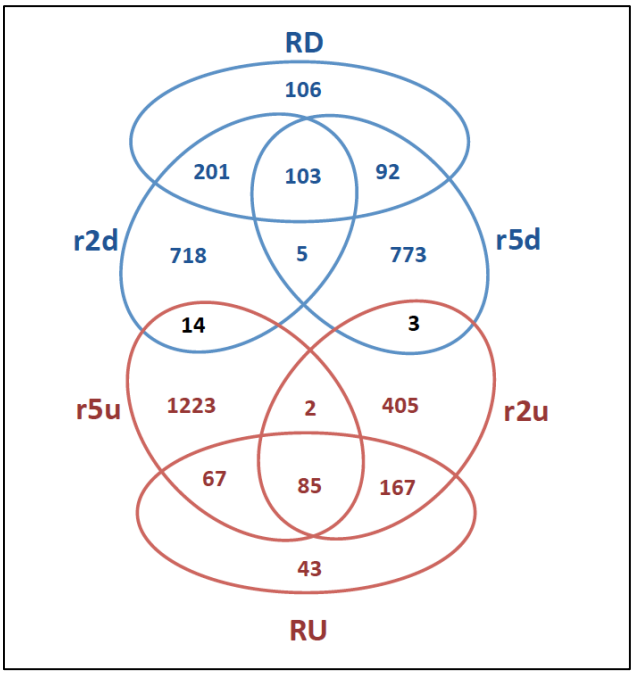


Figure 2.2 Overlap of root differentially expressed transcripts in clone 2 and clone 5. RU: Root both clones combined Upregulated. RD: Root both clones Downregulated. r2u: root clone 2 upregulated. r2d: root clone 2 downregulated. r5u: root clone 5 upregulated. r5d: root clone 5 downregulated.

To further assess the level of variation between clones, we analyzed the 47,117 transcripts with no significant differential expression between control and drought treatment but with substantial expression levels (mean number of reads per base≥5), which we refer to as non-DRTs. We found similar numbers of up- and downregulated non-DRTs in clones 2 and 5 (**Table 2.1**). However, 14,818 non-DRTs showed opposite expression patterns between clones, with 7,335 upregulated transcripts in clone 2 and 7,483 transcripts upregulated in clone 5 (**Figure 2.3**). Of these non-DRTs, 3,455 shared at least a two-fold opposite LFC between clones. As for DRTs, clone 5 exhibited a higher number of genotype-specific transcripts compared to clone 2 (**Table 2.1**).

Altogether, these findings underlie the fundamental difference in the gene expression response to soil dehydration between the two genotypes.

The analysis of DRTs expression level revealed another facet of the divergent response between the two clones. Both up- and downregulated DRTs in clone 5 showed a significantly higher [LFC] than the DRTs in the correspondent expression regimes in clone 2 (upregulated DRTs, Mann-Whitney *U* test, $P = 0$; downregulated DRTs, Mann-Whitney *U* test, $P = 3.55271e\text{-}15$). The distribution of LFC was higher at lower [LFC] in both clones and expression regimes with the exception of the upregulated DRTs in clone 5, which peaked at around LFC=5.5 (**Figure 2.3A-B**). When the DRTs of both clones were combined, the [LFC] was significantly higher in upregulated compared to downregulated transcripts (Mann-Whitney *U* test, $P = 0$). In non-DRTs, [LFC] was also significant more elevated in up- and downregulated transcripts of clone 5 than clone 2 (upregulated DRTs, Mann-Whitney *U* test, $P = 0$; downregulated DRTs, Mann-Whitney *U* test, $P = 0.013$)). Given the distribution of the LFC of non-DRTs (insets in **Figure 2.3A-B**), the significance of these results is likely the product of a high number of data points rather than reflecting a biologically relevant difference in expression levels between non-DRTs of the two clones. Interestingly, the average [LFC] was not significantly different between the 87 upregulated DRTs and the 108 downregulated DRTs shared by clones (Wilcoxon Rank test, $P > 0.05$ for both tests). The LFC distribution of the 87 shared upregulated DRTs mirrored that of the upregulated DRTs of clone 5, with slightly lower central peak around LFC=4.5 in both clones (**Figure 2.3C-D**).

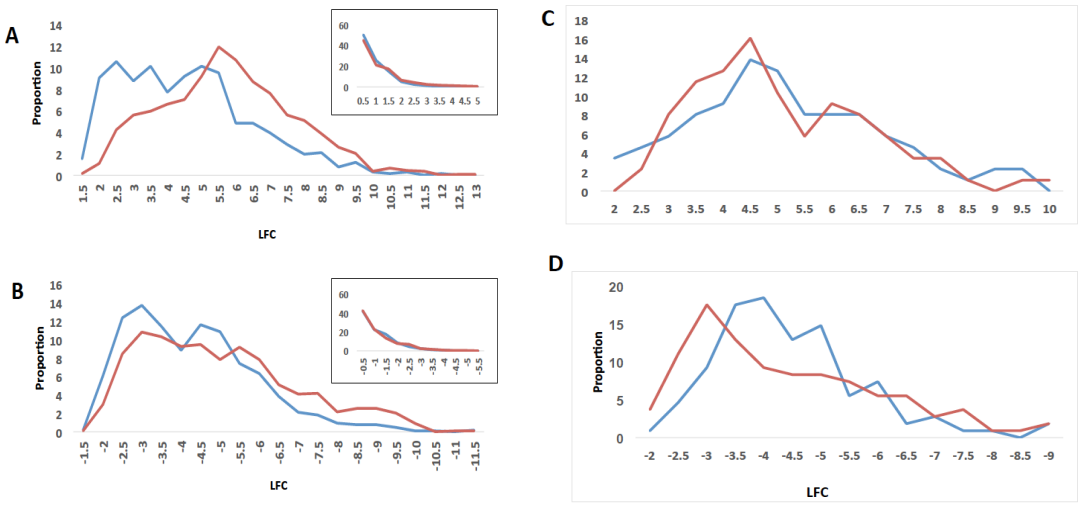


Figure 2.3 Distribution of LFC in clone 2 (red) and clone 5 (blue) between (A) all upregulated and (B) downregulated DRTs, and shared (C) upregulated and (D) downregulated DRTs. The inset in (A) and (B) show the correspondent LFC distributions for non-DRTs.

2.2.5. Functional Annotation of DRTs

We used Blast2GO (GOTZ *et al.* 2008) and EnTAP (HART *et al.* 2020) to functionally annotate the TransDecoder set of transcripts. A total of 48,676 and 38,679 transcripts were functionally annotated by Blast2GO and EnTAP, respectively. Of these, 35,838 where annotated by both programs, with a total of 48,868 transcripts showing evidence of functional annotation. Using the Fisher's test implemented in Blast2GO, we found 190 Gene Ontology categories that were significantly enriched or depleted among clones and expression regimes (up- and downregulated DRTs). A higher number of over- and underrepresented GO terms were found in downregulated DRTs compared to upregulated DRTs (**Figure 2.4**). Depleted GO categories were largely shared across

clones, whereas the few enriched GO terms that overlapped between clones 2 and 5 were found only among downregulated genes (**Figure 2.4**). Enriched GO terms included categories that are expected to be found in drought response experiments, such as "response to water" and "response to abiotic stimulus" in upregulated DRTs in clone 2, and "response to stimulus" in upregulated DRTs in clone 5.

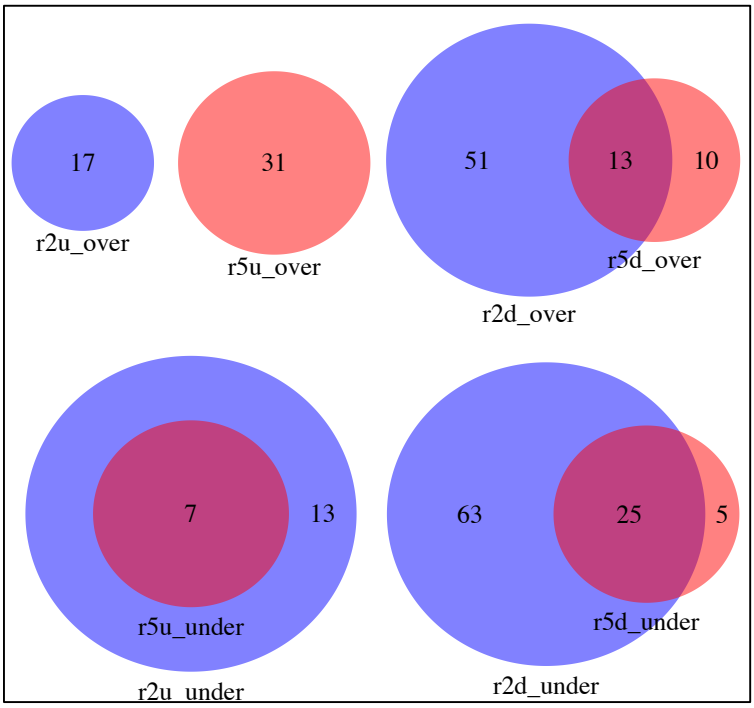


Figure 2.4 GO terms enrichment and depletion between clones and expression regimes. Over: overrepresented GO terms. Under: underrepresented GO terms.

Eighty-seven KEGG pathways were found associated to 293 up- and downregulated DRTs from the two clones. Overall, a higher number of KEGG pathways were found in clone 2 then clone 5, and in downregulated compared to upregulated DRTs (**Table 2.2**). About 45% of KEGG pathways (39/87) were present only in one clone and one

expression regime, but shared pathways were found between most clones and expression regimes, with 7 pathways present in all four types of DRTs (**Figure 2.5**). The number of KEGG pathways showed a weak correlation ($r = 0.38$) with the total number of DRTs in each tested clone by condition. Indeed, only 24 KEGG pathways were represented in the group of 1,391 upregulated DRTs in clone 5, as opposed to the 44 pathways found in 662 upregulated DRTs in clone 2 (**Table 2.2**). This suggests that most DRTs in clone 5, and especially those upregulated in response to drought, are largely not associated with metabolic pathways.

Table 2.2 Total number of DRTs, KEGG pathways, enzymes and DRTs in KEGG metabolic pathways for up- and downregulated DRTs in clone 2, clone 5 and between the two clones

	#Total DRTs	#Pathways	#Enzymes	#DRTs in Pathways
r2d	920	61	46	106
r5d	896	37	29	46
r2u	590	44	34	58
r5u	1,261	24	23	63
RD 2vs5	258	18	15	12
RU 2vs5	281	20	18	20

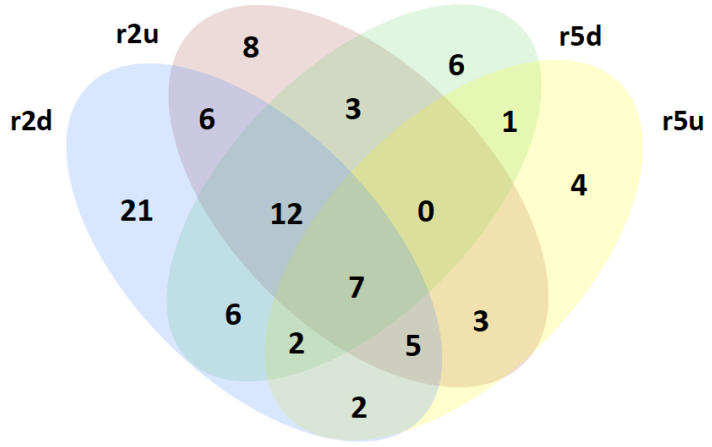


Figure 2.5 KEGG pathways in up-regulated and down-regulated genes of clone 2 and clone5.

In 24 KEGG pathways, DRTs encoded enzymes involved in multiple reactions and thus more likely to represent important metabolic components of the drought response in loblolly. For instance, five reactions were affected by downregulated DRTs in clone 2 in the starch and sucrose metabolism pathway (map00500). Overall, several of these 24 pathways included DRTs across both clones or expression regimes. However, only 28/109 enzymatic reactions and a mere 6/293 DRTs were shared between clones and expression regimes across all KEGG pathways, indicating that different components of the same pathways are often activated in the two clones in response to drought. Overall, we found a few pathways with multiple enzymatic reactions that showed upregulated or downregulated DRTs only. The pathways "Pyruvate metabolism", "Pentose and glucuronate interconversions" and "Thiamine metabolism" contained downregulated DRTs of both clones, whereas several upregulated DRTs in clones 2 and

5 belonged to “Glutathione metabolism”, “Amino sugar and nucleotide sugar metabolism” and “Galactose metabolism” pathways. These metabolic reactions could belong to a core group of pathways activated or repressed in response to drought in loblolly.

To gain further insights into the gene regulatory processes associated with drought tolerance in loblolly we searched for DRTs predicted to encode transcription factors. A total of 1,984 and 1,574 transcripts were predicted transcription factors according to the Blast2GO and EnTAP annotation results, respectively. We also identified 2,110 transcripts with homology to known plant transcription factors using the PlantTFDB (JIN *et al.* 2017). Combining these results on a gene-by-gene basis, we obtained 1,550 predicted loblolly TFs, corresponding to ~4.4% of the 35,220 loblolly genes. All TFs were assigned to families according to the PlantTFDB classification. DRTs included 153 TFs, with fifteen of these DRTs shared between clones (eleven up- and four downregulated genes; **Tables 2.3-2.4**). A higher proportion of TFs was found in upregulated DRTs (3.6-9.5%) compared to downregulated DRTs (2.8-4.4%), and in clone 2 compared to clone 5 (**Table 2.3**). Additionally, more TF families were identified among upregulated than downregulated DRTs (29 vs. 19). Similarly, upregulated DRTs account for most TFs compared to downregulated ones (102 vs. 66, after removing redundant DRTs between clones). Upregulated DRTs from clone 2 showed the highest proportion of TFs, which was driven by a higher than average number of transcripts in multiple families rather than more TF families being present only in this clone and expression regime (**Table 2.3**).

Table 2.3 Number of predicted TFs in all transcripts, both clones and both regimes

	# Genes	TF	% TFs	TF families
All Transcripts	35,220	1,550	4.4	56
non-DRTs	31,858	1,397	4.4	56
DRTs	3,362	153	4.6	30
r2u	598	57 (11)	9.5	20
r5u	1,240	45 (11)	3.6	21
r2d	972	43 (4)	4.4	16
r5d	896	25 (4)	2.8	12

Numbers in parenthesis show shared DRTs between up- or downregulated regimes.

Overall, DRTs encoded TFs from 30/56 families found in the loblolly transcriptome (**Table 2.4**). The percentage of the transcriptome TFs from each family found in DRTs ranged from ~3 to ~38%. Several TF families showed a biased distribution among clones and expression regimes (**Table 2.4**). Of the 30 TF families found in DRTs, only five (bHLH, bZIP, ERF, NAC and RAV) occurred among all clones/regimes, whereas two (NF-YC, Trihelix) were present in both clones upregulated DRTs only, two (Dof and LBD) were found exclusively in clone 2 DRTs and one (MADS) occurred only in clone 5 DRTs. A higher proportion of TFs in the NAC and C3H families was found in upregulated DRTs from both clones, whereas the family WRKY contained mostly downregulated genes. Furthermore, 17 and 10 TF families occurred in upregulated or downregulated DRTs of only one clone, respectively.

Table 2.4 Predicted TFs family in all transcripts, both clones and both regimes

TF family	All genes	non-DRTs	r2u	r5u	r2d	r5d	%DRTs
ARR-B	14	11	3	0	0	0	21.4
bHLH	158	136	6 (1)	3 (1)	6 (1)	9 (1)	15.2
bZIP	56	44	7 (2)	3 (2)	3	1	25.0
C2H2	72	68	0	2	1	0	4.2
C3H	44	40	2 (1)	2 (1)	0	1	11.4
CO-like	11	10	0	1	0	0	9.1
Dof	18	15	2	0	1	0	16.7
ERF	164	149	5	2	5	3	9.1
G2-like	27	25	1 (2)	2 (2)	1	0	14.8
GATA	21	20	0	0	1	0	4.8
GeBP	10	8	0	2	0	0	20.0
GRF	8	6	0	1	1 (1)	1 (1)	37.5
HB-other	10	9	0	1	0	0	10.0
HB-PHD	14	9	0	2	2	1	35.7
HD-ZIP	36	35	1	0	0	0	2.8
LBD	47	43	2	0	2	0	8.5
M-type_MADS	12	11	1	0	0	0	8.3
MADS	60	56	0	3	0	1	6.7
MIKC_MADS	8	7	0	1	0	0	12.5
MYB	165	143	8 (3)	6 (3)	10	0	14.5
MYB_related	75	69	4	1	2	0	9.3
NAC	77	62	7 (1)	6 (1)	2	1	20.8
NF-X1	18	15	1	0	1	1	16.7
NF-YA	7	6	0	1	0	0	14.3
NF-YC	11	9	2 (1)	1 (1)	0	0	27.3
RAV	18	15	1	1	1 (1)	2 (1)	27.8
TALE	12	11	1	0	0	0	8.3
TCP	29	26	1	1	0	2	13.8
Trihelix	61	57	1	3	0	0	6.6
WRKY	64	59	1	0	4 (1)	2 (1)	10.9

Numbers in parenthesis show shared DRTs between up- or downregulated regimes.

To determine whether loblolly DRTs include orthologs to genes known to be involved in drought tolerance in flowering plants, we searched for sequence homology

between DRTs and the 200 genes deposited in DroughtDB, a manually curated database of loci whose role in drought tolerance has been experimentally determined (ALTER *et al.* 2015). We found significant sequence similarity (see **Methods**) between 160 loblolly transcripts from 116 loci and 83 DroughtDB genes (**Table 2.5**). The higher number of loblolly transcripts than DroughtDB genes is due to both the presence of multiple expressed isoforms in some loblolly genes, and to the duplication of some DroughtDB genes in loblolly. Eleven DRTs matched DroughtDB genes. Seven of these DRTs are predicted to be involved in ABA biosynthesis, catabolism or downstream pathways. Nine out of eleven DRTs were upregulated, a significantly higher proportion than downregulated genes (**Table 2.5**; Fisher's exact test, P=0.035). Furthermore, the nine upregulated DRTs exhibited a significantly higher increased in gene expression than all upregulated DRTs combined (**Table 2.5**; Mann-Whitney U test, P=0.0015). Two of these DRTs, MSTRG.33848.1 and MSTRG.57622.1, showed conserved expression patterns in clones 2 and 5. The 149 non-DRTs with homology to DroughtDB genes occurred in both clones with the exception of two transcripts detected only in clone 5. No significant [LFC] differences were found between up- and downregulated transcripts of the two clones. However, Forty-five of these transcripts had opposite expression patterns between clone 2 and clone 5 (**Table 2.5**).

Table 2.5 Loblolly transcripts homology with DroughtDB genes

	DRTs		non-DRTs	
	Upregulated	**Downregulated**	**Upregulated**	**Downregulated**
Clone 2	6	1	65	75
Clone 5	5	1	73	75
Both clones combined (bcc)	6	0	72	75
Clones 2 and 5 opposite	0	0	21	24
Only clone 2	2	1	0	0
Only clone 5	1	1	0	2
Only bcc	0	0	0	0
Only clones 2 and 5	0	0	0	0
Only clone 2 and bcc	2	0	12	14
Only clone 5 and bcc	2	0	15	11
All combined	2	0	43	49
Total	**9**	**2**	**97**	**100**

2.3. Discussions

The genetic basis of drought response variation between different genotypes is poorly understood in conifers. In this study, we performed a transcriptome analysis on root samples of loblolly pine ramets from two clones with different tolerance to water deficit. This represents the first RNA-sequencing investigation in loblolly seedlings grown in drought-simulated conditions, providing more comprehensive gene expression data compared to previous studies based on surveys of a few candidate genes, or ESTs/microarray data.

We found that the vast majority of DRTs exhibit a GxE pattern of expression in the two clones. Strong GxE effects were observed especially at the level of individual genes, with very little overlap of upregulated and downregulated genes between the two clones. Although the direction of expression change was largely the same in genes

between clones, twice as many upregulated genes under drought stress were found in the more drought tolerant clone (clone 5), suggesting that increased drought tolerance in some loblolly genotypes is associated with the ability to activate a larger group of genes compared to drought-sensitive genotypes. Approximately 20% of DRTs (819/4,012) showed an opposite expression pattern between clones, including many transcripts with significant differential expression only in one clone. Furthermore, both up- and downregulated DRTs in clone 5 showed significantly higher absolute $\log_2$ fold change (LFC) compared to those of clone 2. Extensive GxE effects were also observed in the 47,117 non-DRTs, with 14,818 transcripts showing opposite expression patterns between clones and 1,053 transcripts present only in clone 2 or clone 5.

The GxE pattern was less pronounced at the level of predicted gene functional categories or metabolic pathways. Indeed, the gene ontology and metabolic pathways enrichment analyses indicate that similar functional groups of transcripts are differentially expressed under water stress in clone 2 and clone 5. However, upregulated DRTs showed no shared biological processes between the two clones. Altogether, these findings lend support to the notion that water deficiency elicits a response based on remarkably different genes and genetic networks at the root level in the two loblolly genotypes examined here. This conclusion is further supported by the analysis of differentially expressed transcription factors. Only 11/102 upregulated TFs and 4/66 downregulated TFs were shared between the two clones. Furthermore, a similar number of TF families were shared between clones and expression regimes, and many TF

families in up- or downregulated DRTs occurred only in one clone. Thus, very few TFs and TF families were shared between the two clones.

These results are in contrast with a previous microarray-based analysis showing remarkable similarities in the gene expression patterns between drought-stressed, well-watered and drought-recovered treatments in roots across 4 loblolly clones (LORENZ *et al.* 2011). The different genotypes, treatment regimes and gene expression detection technologies between our study and Lorenz and collaborators' work may all contribute to these discrepancies. Notably, low levels of GxE have also been reported in the root transcriptome of different genotypes exposed to drought stress among flowering plant species. For instance, the wheat tolerant cultivar JM-262 and susceptible cultivar LM-2 showed largely overlapping sets of both up- and downregulated DRTs (HU *et al.* 2018). In a different study, four wheat varieties showed on average a 51% overlap between root DRTs (MIA *et al.* 2020). High levels of congruence between DRTs of drought tolerant and sensitive genotypes/cultivars have also been reported in rice (BALDONI *et al.* 2016; LOU *et al.* 2017), barley (JANIAK *et al.* 2018), maize (ZHANG *et al.* 2019) and poplar (COHEN *et al.* 2010). Although genotypes with varying drought tolerance clearly show remarkable differences in the gene expression response during water deficiency, these differences appear to be especially pronounced between loblolly clones 2 and 5. We recognize that our conclusions might have been affected by some caveats. Both significantly enriched or depleted functional categories and metabolic pathways contained a relatively small proportion of DRTs. Thus, clones 2 and 5 could share a higher proportion of functional groups and metabolic network that showed by our

analyses of GO terms and KEGG pathways. Furthermore, we applied a prolonged drought treatment that mimic more closely the water deficiency regimes experienced by loblolly pine forests, which might elicit a different genetic response compared to analogous experiments that largely test "acute" drought conditions enforced for a short period of time. Further studies are warranted to determine if and how the gene expression profile changes between acute and prolonged drought treatments in loblolly genotypes.

Clone-specific genetic networks involved in abiotic stress responses can be activated or repressed by modified expression of key transcription factors. Therefore, we prioritized the identification of differences in TFs expression between clones and treatments. Transcripts encoding for transcription factors from a variety of families were identified among DRTs. Many of these TFs are known to be expressed in response to drought, including the dehydration response element binding factors (DREBs) of the ethylene responsive factor (ERF) family (XIE *et al.* 2019), the ABA response elements (ABREs) of the basic leucine zipper (bZIP) domain family (GOLLDACK *et al.* 2014), and TFs from the WRKY (TRIPATHI *et al.* 2014), NAC (NURUZZAMAN *et al.* 2013) and MYB (BALDONI *et al.* 2015) families. Similar cohorts of TF families were identified in drought-response gene expression experiments in conifers, including loblolly pine (LORENZ *et al.* 2011), as well as in flowering plants (JANIAK *et al.* 2016). We further identified several TF families that have been increasingly recognized in association with drought and may play a major role in the response to water deficit in loblolly. Trihelix TFs, which include the GT factors, are present among upregulated DRTs but do not

appear in downregulated DRTs. Some Trihelix TFs are expressed in response to abiotic stress, including drought, in multiple angiosperms (XIE *et al.* 2009; MU *et al.* 2016; YU *et al.* 2018; MAGWANGA *et al.* 2019). The largest group of TFs in our dataset is represented by the basic helix-loop-helix (bHLH) family, which includes several up- and downregulated DRTs from both clones. This family alone is suggestive of the complexity of the regulatory networks involved in the response to drought and similar abiotic stressors in loblolly; among the 23 DRTs encoding a bHLH TF, only 1 downregulated DRTs was shared between clones. In agreement with previous studies in conifers, we found that most TFs whose expression changed significantly in response to drought were upregulated. Nevertheless, we observed an elevated number of downregulated TFs in our experiments compared to the microarray results of Lorenz et al. (2011), even though these authors found more downregulated than upregulated DRTs. This implies that the downregulation of TFs may play a more important role than previously recognized in the root drought response of loblolly.

Among the 200 experimentally identified drought-related genes reported in DroughtDB, we identified 83 with high homology with one or multiple loblolly transcripts. Given the relatively stringent thresholds we applied to detect homology, it is likely that more known DroughtDB genes are present in loblolly. Additionally, some DroughtDB genes are likely to be not expressed in root tissues. The finding that upregulated DRTs with homology to DroughtDB genes are expressed at higher levels than other upregulated DRTs suggests that this small group of genes might play a critical role in drought response. This is further supported by the fact that six of these genes are

involved in ABA biosynthesis, catabolism or downstream pathways. The role of other DroughtDB genes expressed in loblolly in response to aridity is less clear, especially those showing opposite expression patterns between clones. This suggests that while some genes might share a key function in drought response in both angiosperms and loblolly, many components of the genetic networks activated and repressed in low water availability conditions could differ between flowering plants and gymnosperms.

2.4. Materials and Methods

2.4.1. Plant materials and experimental design

The loblolly pine varieties were provided by ArborGen Inc. A total of 140 ramets (20 for each variety) were planted on September 25, 2014 in a greenhouse operated by the Department of Ecosystem Science and Management at Texas A&M University in College Station, TX. After four weeks of growth in well-watered conditions, ramets of each variety were randomly assigned to 5 blocks (replicates) for each of two treatments, well-watered (control) and low-watered (drought-simulated), which were watered 1 out of every 6 times the control ramets were. Two drying periods were applied, from December 2014 to March 2015 and from mid-April 2015 to the end of May 2015. All ramets were grown in sand with periodic fertilizer addition, with automatic watering adjusted based on soil moisture and pre-dawn water potential measurements. Ramets from the three varieties 2, 5 and 6 were selected for further growth and sampling based on gas exchange preliminary data taken in December 2014 showing differences between varieties in stomatal conductance and photosynthetic rate. After six months, the varieties

2 and 5 showed the highest difference in water potential and were selected for phenotype and transcriptome (RNA sequencing) analyses. Six ramets from each of the two varieties (three ramets per treatment) were harvested on the morning of May 29, 2015. Harvested tissues to be used for transcriptome analyses were wrapped by marked aluminium foil paper and immediately stored in an -80°C freezer.

2.4.2. Physiological measurement and treatment comparison

Physiological and other phenotypic measurements including growth estimate, soluble carbohydrate, $\delta^{13}C$, pre-dawn and mid-day water potential, gas exchange measurements, specific leaf area and leaf nitrogen content were carried out on these ramets.

2.4.3. RNA extraction and cDNA sequencing

Total RNA was extracted from whole needles and part of the root system (~100 mg each) for each harvested ramet. The RNA was isolated after grinding each sample in liquid nitrogen. RNA samples with a RQN, which is RNA Quality Number, between 5.2 and 10.0 were used for RNA-sequencing (RNA-seq) experiments. Quality control, library preparation, sequencing and preliminary data filtering were performed by the Texas A&M AgriLife Genomics and Bioinformatics Services. RNA-Seq libraries were constructed using the Illumina TruSeq RNA Sample Preparation Kit, as per manufacturer instructions. cDNA libraries were sequenced using the high-throughput RNA-Seq technology. All libraries were quality checked and sequenced on two lanes of

Illumina HiSeq-2500 platform using a 2x125bp paired-end strategy. One needle library contained mostly bacterial DNA and was thus removed from downstream analyses. Sequencing of the twenty-three remaining samples generated 568.2 million raw reads (~120 Gb) reduced to 514.6 million reads after pre-filtering (see below). The average reads number was 24,992,695 and 19,518,450 for each root and needle library, respectively.

2.4.4. Reads data filtering

More than 95% of de-multiplexed reads passed the instrument-level pre-filtering and were further processed. The pre-filtered reads were checked using FastQC (ANDREWS 2010). Filtering was applied to the raw data to generate clean reads with the following approach. First, the program SortMeRNA (KOPYLOVA *et al.* 2012) was used to identify and remove reads corresponding to rRNA genes. On average, 4.24% of reads were removed from each library in this step. Second, adapters were cut from the reads allowing maximally 2 mismatches under the quality score threshold 30 using Trimmomatic version 0.35 (BOLGER *et al.* 2014). Reads were scanned with a 4-base wide sliding window and cut when the average quality per base dropped below 14, and reads with less than 50 bases long after the trimming steps were dropped. Finally, we implemented a stringent filtering process after mapping reads onto the genome assembly v 1.01 in order to account for the high level of sequence redundancy in the large loblolly pine genome. Cleaned reads from the previous two steps were aligned to the loblolly pine genome v 1.01 using HiSAT2 (KIM *et al.* 2015), applying default parameters except

min-intronlen and max-intronlen set to 30 and 10000000, respectively. Subsequently, we removed reads that do not map concordantly on a single locus or have >3 mismatches by retaining only reads with the following parameters in the SAM output: NH:i:1, YT:Z:CP and XM:i:0-3. This step allowed reducing the mapping of reads to incorrect loci.

2.4.5. Transcriptome assemblies

An overall transcriptome was first built with all the clean reads using StringTie (PERTEA *et al.* 2015), which assembles and quantifies the transcripts including novel splice variants in each library. A second assembly was then generated using the Stringtie merge function to construct one set of transcripts, which was consistent across all 46 samples with better read coverage. Candidate coding regions were retrieved using TransDecoder (https://github.com/TransDecoder) based on merged transcript sequences. The transcripts abundances for each library were re-computed by StringTie based on the newly constructed candidate coding transcriptomic structure. The filtered high-quality reads were assembled and merged by Stringtie to get a total number of 54,826 transcripts with an N50 length of 1,440 bp. The re-estimation from the assembly results of each library against the merged transcriptomic data was carried out, resulting in transcripts expression value count matrix.

2.4.6. Genetic distance

SNPs between each library and the loblolly assembly v1.01 reference sequence were detected using the programs Opossum and Platypus (OIKKONEN AND LISE 2017).

Opossum was used to pre-process the assembled data for each library, whereas variant-detection calling was carried out with Platypus using reads realignment to the genome assembly to achieve both high sensitivity and high specificity. Candidate variants were filtered based on PASS and Quality of 100 or above, and then the ones supported by a minimum of 10 reads coverage were kept by Platypus. Genetic distances were calculated as the number of SNPs divided by the total number of aligned nucleotides between each library and the genome assembly.

2.4.7. Quantitative qRT-PCR

Twenty-six transcripts with varying degrees of differential expression between control and drought-stressed ramets were selected for qRT-PCR experiments. Twenty-two out of twenty-six transcripts were used in qRT-PCR analysis based on their primer design process results. All the primers of the twenty-two selected transcripts were passed with a length cutoff between 21 to 27 base pairs, an E-value smaller than 2e-04 and a score greater than 41. The qRT-PCRs were performed using the SYBR kit on an ABI 7500 Real-Time System (Applied Biosystems). The Actin unigene and 98 unigene were used as internal controls to normalize the expression values based on their consistent expression level across tissues. The relative quantitative method ($\Delta\Delta$CT) was used to calculate the fold change in the expression levels of target genes. All reactions were performed in three technical replicates using two biological samples.

2.4.8. Gene differential expression identification

Gene expression values were calculated for each library using Fragments Per Kilobase of transcript per Million mapped reads (FPKM). A final clean transcripts count matrix was applied to the statistical package DESeq2 (LOVE *et al.* 2014), which provided negative binomial generalized linear models to test differential expression across treatments, tissues and clones. Transcripts differential expression was conducted by the DESeq2 count matrix input protocol using collapsing technical replicates function and took other factors as background when comparing two levels in one specific factor. The P-value for each differentially expressed transcript (DET) was adjusted using the Benjamini and Hochberg's approach for controlling the false discovery rate (COLQUHOUN 2014). The moderated log fold changes proposed by Love, Huber, and Anders (2014) used a normal prior distribution, centered on zero and with a $\log_2$ scale, which has been normalized with respect to library size that is fit to the data. In this study, transcripts with an FDR < 0.05 and absolute log2 fold change ≥ 1 were considered differentially expressed.

2.4.9. Gene Annotation and Network analysis

Functional annotation of transcripts was performed using the Blast2GO Professional suites (GOTZ *et al.* 2008). All the transcripts sequences were queried against the NCBI database using Blast and InterProScan default Blast2GO settings, and the results of the two searches were merged in a single annotation output. For functional annotation, Gene Ontology terms were retrieved according to Blast hits for each transcript by mapping and annotation. GO enrichment analysis was performed on the

annotated sequences to show the abundant and scarce GO terms in upregulated and downregulated DRTs in each clone compared to the whole set of transcripts. The Fisher's exact test and Gene Set Enrichment Analysis (GSEA) were conducted for the enrichment analysis. KEGG pathways maps were then extracted through enzyme code mapping of functional annotation in Blast2GO. EnTAP (Eukaryotic Non-Model Transcriptome Annotation Pipeline) (HART *et al.* 2020) was also applied to all the transcriptome transcripts and the corresponding annotation were retrieved.

Transcription factors were annotated by searching the PlantTFDB (JIN *et al.* 2017) using the protein sequences of all transcripts obtained with TransDecoder. The Blast2GO and EnTAP annotation results were searched for TF family names from the PlantTFDB classification scheme and for the key words DNA-binding, DNA binding, transcription factor, regulation of gene expression, regulation of transcription. The annotation entry of retrieved transcripts encoding TF but with no obvious affiliation to a specific TF family were further inspected to identify gene symbols associated with families, i.e. DREB, which belongs to the ERF family. Gene symbols of matching genes from *Arabidopsis thaliana* were searched on the TAIR database (BERARDINI *et al.* 2015). Protein sequences of some transcripts were used in some instances to find corresponding TFs through sequence similarity searches against proteins on the NCBI-BLAST nr database (JOHNSON *et al.* 2008).

Protein sequences of genes deposited in the DroughtDB (ALTER *et al.* 2015) were retrieved from the TAIR10 gene set (BERARDINI *et al.* 2015) when present in *A. thaliana* or from DroughtDB itself. Homologous genes to these sequences were searched among

the TransDecoder set of ~60,000 transcripts from this study using a tBlastn local search approach (CAMACHO *et al.* 2009). The Blast results were parsed with an in-house perl script. Transcripts with at least 60% sequence identity over more than half the length of drought genes were considered homologous sequences. Transcripts with 50-60% sequence identity with drought genes but with alignments containing 10% or more gaps were also considered homologous sequences. In transcripts with homology with multiple entries in DroughtDB, only the blast hit with the highest sequence percentage identity was retained.

2.5. Conclusions

We have found that the root transcriptomic response to water deficiency between tolerant and sensitive loblolly pine clones exhibits a strong GxE pattern across more than 50,000 expressed transcripts. Most up-and downregulated drought-related transcripts, or DRTs, and their expression levels, differed markedly between the two clones. Similarly, we observed limited overlap between metabolic pathways, functional gene categories and transcription factors associated with DRTs between the two clones. These findings suggest that a prolonged water deficit in the roots of different loblolly genotypes elicits genetic responses that diverge beyond what has been observed between drought tolerant and sensitive genotypes in flowering plants, and in previous studies in loblolly. Further studies in *Pinus taeda* and other conifers are warranted to determine the extent of this expression divergence between genotypes across this group of gymnosperms. Linking

the observed divergence to local adaptation in loblolly should also be a major goal of future works in this species.

3. EVOLUTIONARY CONSERVATION OF TRANSCRIPTION FACTOR BINDING SITES IN DROUGHT-RELATED GENES OF LOBLOLLY PINE AND ANGIOSPERMS

3.1. Introduction

The promoter regions of eukaryotic genes contain a variety of transcription factor binding sites (TFBSs), short (5-15 bp) DNA sequences regulating the timing, tissue specificity and duration of transcription (FICKETT AND HATZIGEORGIOU 1997; JUVEN-GERSHON *et al.* 2008). TFBSs in the promoter region constitute major components of a gene's *cis*-regulatory apparatus, which also consist of enhancers, insulators, silencers and other short regulatory sequences (WITTKOPP AND KALAY 2012). A large body of literature suggests that changes in *cis*-regulatory elements (CREs), particularly TFBSs, are responsible for most of the divergence in gene expression patterns between species (REVIEWED IN SIGNOR AND NUZHDIN 2018). Concurrently, many homologous genes are expected to maintain similar expression levels and breadth across distantly related species due to their role in 'housekeeping' cellular processes that remain largely unaltered across the tree of life. Moreover, even slight changes in the short sequence of TFBSs may dramatically affect their ability to bind transcription factors, thus constraining the evolutionary dynamics of these regulatory elements. Thus, sequence conservation of a fraction of TFBSs should be expected even between distantly related organisms. Indeed, several such examples have been described in the so-called 'core promoter', a DNA region that encompass the transcription start site of most eukaryotic

genes and often include the TATA box, Initiator (Inr), downstream promoter element (DPE), motif ten element (MTE) and polypyrimidine initiator (TCT) (JUVEN-GERSHON *et al.* 2008; DANINO *et al.* 2015; ROY AND SINGER 2015). However, the evolutionary conservation of most TFBSs that are critical to the spatial and temporal expression pattern of genes is poorly understood.

The availability of genomic and transcriptomic data from a variety of species could theoretically provide the sources of estimation of the proportion of TFBSs that are conserved between two given species. Because large-scale datasets of experimentally validated TFBSs remain unavailable in most species, genome-wide surveys of *cis*-regulatory elements are primarily conducted using bioinformatic approaches, which have shown relatively high levels of sensitivity and specificity (BAILEY *et al.* 2009). Methods that identify putative motifs *de novo* are especially intriguing because they allow to help collect information without *a priori* inferences on the sequences of regulatory elements. Hence, these methods can in theory produce collection of both conserved and lineage-specific TFBSs. Importantly, these sequences can then be compared to extensive datasets of known TFBSs to determine their homology and identify evolutionary conserved elements (HIGO *et al.* 1999).

In animals, these approaches have revealed that some *cis*-regulatory elements are conserved across distantly related vertebrates (MAESO *et al.* 2013), although in general TFBSs are less conserved in vertebrates than in *Drosophila* species (VILLAR *et al.* 2014). Among plants, genome-wide analyses of conserved noncoding DNA regions have revealed instances of conservation in promoter regions and motifs, at least among

grasses (TURCO *et al.* 2013; BURGESS AND FREELING 2014). A number of other studies have dissected the evolutionary dynamic of promoters in individual genes. For example, the sequence of the *A. thaliana* root hair–specific *cis*-elements (RHEs) upstream of the expansinA7 gene (At *EXPA7*) was retrieved in the promoter region of *EXPA7* orthologs in several other flowering plants (KIM *et al.* 2006). Similarly, the promoter region of the key regulator of plant circadian clock LATE ELONGATED HYPOCOTYL (LHY) contains both a G-box motif (CACGTG) and a 5A motif (AAAAA) that is conserved among orthologous *LHY* genes of *Arabidopsis thaliana*, grapevine and poplar (SPENSLEY *et al.* 2009). Other examples included promoter motifs upstream of genes involved in the jasmonic acid pathway that are shared between *A. thaliana*, *Brassica rapa*, poplar and grapevine (HICKMAN *et al.* 2017) and octamer motifs identified upstream of time-of-day transcriptional networks genes in *A. thaliana*, poplar and rice (MICHAEL *et al.* 2008). Broader studies have provided evidence of widespread conservation of part of the sequence of some binding sites. Analyzing hexamer and octamer motifs occurring at high frequency in promoter regions of *A. thaliana* and rice, Yamamoto et al. (YAMAMOTO *et al.* 2007) found that about ~40% (283/715) of motifs were shared between the two species. Comparing genes from co-expression networks in *A. thaliana* with their poplar orthologs, Vandepoele et al. (VANDEPOELE *et al.* 2009) identified 866 non-redundant 8-mer motifs, 63% of which corresponded to known plant TFBSs. In a recent study, the core sequence of short response elements (REs) recognized by the TF families auxin response factor (ARF) and abscisic acid response elements (ABRE)

biding factors were found to be conserved among hundreds of orthologous genes across 45 eudicots and monocots (LIEBERMAN-LAZAROVICH *et al.* 2019).

Overall, these studies suggest that a significant proportion of TFBSs are shared across distantly related flowering plants. Conversely, it remains unclear to what extent TFBSs are conserved across land plants, largely because of the lack of genomic resources beside the angiosperm lineage. One of the few examples of TFBS conservation across land plants is represented by the ABRE motif, which has been found to share high level of sequence similarity between angiosperms and the moss *Physcomitrella patens* (TIMMERHAUS *et al.* 2011). Other studies have pointed to some level of sequence conservation between flowering plants and gymnosperms, two sister lineages that separated approximately 300 million years ago (BOWE *et al.* 2000). In most cases, these works have characterized TFBSs that are known to be involved in the response to drought and other environmental stresses. The first gymnosperm TFBS was characterized by Loopstra and collaborators in loblolly pine (*Pinus taeda* L.) and consisted of a 7 bp sequence in upstream of both PtX3H6 and PtX14A9 genes; this motif shared high sequence similarity with a TFBS regulating the vascular-specific expression of glycine-rich protein in the common bean (LOOPSTRA AND SEDEROFF 1995). Other regulatory motifs have been subsequently identified in a variety of gymnosperms. For example, in *Pinus sylvestris*, the transcription factor binding sites bHLH and bZIP are homologous to TFBSs found in *Populus trichocarpa* and *Arabidopsis thaliana*. These genes modulate the abscisic acid (ABA) and gibberellic acid (GA) response in drought conditions (VUOSKU *et al.* 2018).

Similarly, the DNA sequencing and analysis of the putative promoter region of the gene BABY BOOM2, encoding a member of the APETALA2/ETHYLENE RESPONSE FACTOR (AP2/ERF) family of transcription factors, in the larch hybrid *Larix kaempferi*×*L. olgensis* have shown multiple sequences with similarity with angiosperm TFBSs (WANG *et al.* 2019). Putative TFBSs of this gene include the dehydration and dark response element ACGT, the DREBP (abiotic stress) regulatory element RYCGAC, the motifs ACGTSSSC, ACACNNG and ACCGAC involved in abscisic acid (ABA) responsiveness, and the motif TTGAC, which is recognized by WRKY transcription factors (IMIN *et al.* 2007; RIGAL *et al.* 2012; WANG *et al.* 2019). In Douglas-fir (*Pseudotsuga menziesii*), a motif in the sequence of the luminal binding protein (BiP) gene promoter sharing sequence conservation with angiosperms regulatory motifs has been found upstream of the heath-shock response gene *HSP70*. This sequence includes an AT-rich *cis*-acting regulatory domain 1 (CRD1) and a second activating domain (CRD2), which binds to the BiP promoter (BUZELI *et al.* 2002; YEVTUSHENKO AND MISRA 2018).

In the Cupressaceae genus *Taxus*, genes involved in the biosynthesis of the antitumorigenic molecule taxol have been intensely characterized. These efforts have led to the discovery of multiple TFBSs that are shared with angiosperm genes (BUZELI *et al.* 2002; YEVTUSHENKO AND MISRA 2018). Notably, some of these genes are also involved in the *Taxus* response to environmental stresses. For instance, the G-box CRE characterized in the taxane 5α-hydroxylase in *Taxus baccata* is bound by the TcMYC transcription factor responsible for up-regulating the expression of multiple genes in

response to drought and high-salinity stresses (YANFANG *et al.* 2018). Other regulatory motifs implicated in the ABA-dependent response to drought have been discovered in *Taxus* cell cultures (SANCHEZ-MUNOZ *et al.* 2018) and in the promoters of CYP450 genes of *Taxus chinensis* (AMBAWAT *et al.* 2013; LIAO *et al.* 2017).

The *PpNAC1* transcription factor plays a critical role in regulating the phenylalanine biosynthesis pathway in conifers. Computational analyses of the *PpNAC1* putative promoter region in maritime pine (*Pinus pinaster*) revealed six SNBEs (secondary wall NAC binding element) and one AC element of the AC-II (ACCAACC) class. Using electrophoretic mobility shift assays, Pascual et al. found that *PpNAC1* is self-regulated through the interaction of PpNAC1 with the SNBE motifs (PASCUAL *et al.* 2018). In another study, the spermidine synthase gene in Scots pine (*Pinus sylvestris*) and loblolly pine has been found to contain several sequences with similarity to known angiosperm TFBSs (VUOSKU *et al.* 2018).

Although broad bioinformatic analyses of gymnosperm promoter regions have not been carried out, a recent study has investigated the regulatory landscape of the large family of dehydrin genes across both angiosperm and gymnosperm genomes. Dehydrins are proteins involved in the response and adaptation to abiotic stress in plant development (ZOLOTAROV AND STROMVIK 2015). In this work, 350 dehydrin promoter sequences from 51 plants including Norway spruce and loblolly pine were analyzed to computationally identify regulatory motif *de novo*. Dehydrins were separated in five classes based on the occurrence of specific amino acid segments in their protein sequences. Loblolly and Norway spruce dehydrins were found in the two classes K_n and

SK_n. A total of nine discovered TFBSs were identified in these two classes and presumed conserved across land plants; however, this study did not reveal if these motifs were detected in the two conifer genomes. Therefore, the extent of TFBS conservation between dehydrins in angiosperms and gymnosperms remain unclear.

While important, these studies have focused on promoter regions from single genes, pathways or gene families. Additionally, most investigations on gymnosperm TFBSs focused on detecting known angiosperm motifs and were not suited to identify possible gymnosperm-specific regulatory motifs. Therefore, the overall evolutionary conservation between angiosperms and gymnosperms TFBSs remains unknown. These fundamental aspects of gene expression regulation can now be addressed leveraging on the recent sequencing and annotation of multiple gymnosperm genomes (BIROL *et al.* 2013; NYSTEDT *et al.* 2013; WEGRZYN *et al.* 2014; WARREN *et al.* 2015; GONZALEZ-IBEAS *et al.* 2016; GUAN *et al.* 2016; NEALE *et al.* 2017; WAN *et al.* 2018; MOSCA *et al.* 2019). These resources have prompted a number of comparative genomics studies that have shown a low rate of DNA sequence evolution in this group compared to angiosperms (De La Torre et. al 2017) contrasting with a rapid pace of gene duplication and loss, at least in Pinaceae (NEALE *et al.* 2017; CASOLA AND KORALEWSKI 2018). However, little is known about the evolutionary dynamics of promoter regions in gymnosperms and their level of sequence and functional conservation with angiosperms' promoters, especially at a genome-wide scale.

In this work, we performed the first large-scale *de novo* survey of promoter motifs in gymnosperms, using a dataset of more than 4,000 transcripts from 3,495 genes

that are differentially expressed in response to drought across two loblolly clones. We identified thousands of putative regulatory sites corresponding to 24 non-redundant TFBSs that show significant sequence homology with known regulatory motifs of angiosperms. Our results suggest that seed plants experience a high level of promoter motifs conservation in genes implicated in the response to abiotic stress.

3.2. Results

3.2.1. Discovery of TFBSs in putative promoter regions of loblolly pine drought-related genes and analysis of their conservation in angiosperms

In a recent study, we identified 4,012 drought-related transcripts (DRTs), corresponding to 3,495 unique loci, by comparing the gene expression profile of control and drought-stressed root seedlings in two loblolly genotypes, named hereafter clone 2 and clone 5, using RNA-sequencing data (Li et al., *unpublished*). Taking advantage of the loblolly whole-genome assembly, we retrieved 24,794 and 15,611 putative promoter regions from 1,100 bp and 2,100bp upstream of 60,090 transcripts, respectively, using a stringent set of criteria to limit the number of possible false positives (see Methods). We identified 1,478 DRTs with promoter regions. Depending on their expression pattern in response to drought and genotypes, we assigned these DRTs to four datasets, namely **r2u** (root clone 2 upregulated), **r2d** (root clone 2 downregulated), **r5u** (root clone 5 upregulated) and **r5d** (root clone 5 downregulated) (**Table 3.1**). The longer putative promoter regions contained ~60% of MEME and Seeder motifs.

Table 3.1 Drought related transcripts promoter datasets and TFBSs.

DRT Datasets	Promoter Length (bp)	Number of DRTs	#Sites	#Unique TFBSs
r2u	1,100	323	2,157	20
	2,100	200		
r2d	1,100	519	4,881	15
	2,100	320		
r5u	1,100	523	2,157	4
	2,100	292		
r5d	1,100	378	299	3
	2,100	236		

r2u: root-clone2-upregulated DRTs. r2d: root-clone2-downregulated DRTs.
r5u: root-clone5-upregulated DRTs. r5d: root-clone5-downregulated DRTs.

To detect TFBSs, we retrieved DNA sequences corresponding to proximal (1,100 bp) and distal (2,100 bp) upstream regions of the putative Transcription Start Site (TSS) of the four DRT datasets. Only DRTs with an annotated 5'UTR were used in this analysis. Using the programs MEME (BAILEY *et al.* 2009) and Seeder (FAUTEUX *et al.* 2008), we identified 9,494 motif sites corresponding to 42 unique motif sequences in 1,386 DRTs (**Table 3.1**). A total of 1,096 sites for 17 unique motifs were retrieved by MEME, compared to 8,468 sites and 25 unique motifs obtained by Seeder. The number of sites per TFBS ranged from 7 in the MEME YGGCCGTCRR motif to 857 in the Seeder CGCGTGTA TFBS. On average, 54 and 368 sites were found for MEME and Seeder motifs, respectively. Downregulated DRTs from clone 5 showed approximately an order of magnitude fewer TFBSs then the other three datasets, despite a comparable number of putative promoter sequences (**Table 3.1**). Additionally, Clone 2 datasets were characterized by a several-fold higher number of unique TFBSs than clone 5 (**Table**

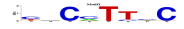
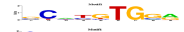
3.1). The number of unique TFBSs was also higher in clone 2 vs. clone 5 datasets (**Tables 3.1-3.2**).

Table 3.2 Summary of TFBSs by DRT datasets.

DRT Datasets	TFBSs	Program	#Sites
r2u	AAAAHAAAAA	MEME	132
r2u	CAMGTGGCGG	MEME	34
r2u	CCACKTGTCG	MEME	42
r2u	CCCAATTGAC	MEME	24
r2u	CCCCYBCCCC	MEME	79
r2u	CGGCCAAATC	MEME	42
r2u	CGGCCGTCAA	MEME	39
r2u	GATCTGGCCG	MEME	28
r2u	GCCACGTGTC	MEME	77
r2u	GGACGGCCMR	MEME	9
r2u	TCGCGCATCC	MEME	33
r2u	TGGACGGCCC	MEME	17
r2u	YGGCCGTCRR	MEME	7
r2u	YYGACGGCCR	MEME	15
r2u	ACGTGGCG	Seeder	294
r2u	ACGTGGCT	Seeder	144
r2u	ATCGTTCG	Seeder	191
r2u	CGACGGCG	Seeder	359
r2u	CGGATAAG	Seeder	285
r2u	CTACGTGT	Seeder	306
r2d	ATTTTTWTTT	MEME	255
r2d	CAACCWSCCA	MEME	83
r2d	GCCACCMCCR	MEME	36
r2d	ACGTAATA	Seeder	308
r2d	ATATATAA	Seeder	454
r2d	CGCGCACG	Seeder	493
r2d	CGCGTGCG	Seeder	289
r2d	CGCGTTAC	Seeder	491
r2d	CGTCGTTA	Seeder	314
r2d	GCGTCGTA	Seeder	511
r2d	TAAATATA	Seeder	262
r2d	TAAATTAA	Seeder	280
r2d	TACATATA	Seeder	308
r2d	TCGCGCGT	Seeder	505
r2d	TTAATTAA	Seeder	292
r5u	AATACGCG	Seeder	512
r5u	ATACGCGT	Seeder	510
r5u	CGCGCCGC	Seeder	278
r5u	CGCGTGTA	Seeder	857
r5d	GAGSCTCCMA	MEME	37
r5d	GGGCGGGTGC	MEME	37
r5d	AATAATAT	Seeder	225

The program STAMP (MAHONY AND BENOS 2007) was then applied to determine if the predicted loblolly TFBSs were functionally related to known angiosperm motifs deposited in the Plant *cis*-acting regulatory DNA elements (PLACE) database (HIGO *et al.* 1999). All the 42 TFBSs showed a significant match with angiosperm TFBSs. Combining the sequence information of each motif and the STAMP results, we obtained 24 functionally homologous TFBSs, five of which were identified by both MEME and Seeder (**Table 3.3**). The following analyses focus on these TFBSs. Eleven of the 24 TFBSs were associated to motifs known to be involved in abiotic stress responses in plants (**Table 3.3**).

Table 3.3 Features of DRTs *cis*-regulatory elements in loblolly pine

Dataset	Upstream region	Motif	STAMP Match	E-value	WebLogo	Programs	Description	Number of DRTs
r2d	1100	CAACC[TA][CG]C[CA]A	PROXBBNNAPA	2.11E-05		M	Required for seed specific expression and ABA responsiveness	74
r2d	2100	ATTT[TA]T[TA]TTT	MARTBOX	6.32E-10		M,S	Conserved drought responsive elements in gymnosperm and angiosperms	198
r2d	2100	GCCACC[AC]CC[GA]	SBOXATRBCS	4.87E-05		M,S	Important for the sugar and ABA responsiveness	27
r2d	1100	CGCGTTAC	AMMORESIIUDCRNIA1	2.27E-05		S	Involved in ammonium-response	491
r2d	2100	CGCGTGCG	MYCATERD1	8.48E-08		S	Necessary for expression of erd1in dehydrated Arabidopsis	315
r2d	2100	CGTCGTTA	RBCSBOX3PS	9.72E-08		S	Bind the transcription factor GT-1	314
r2d	2100	TACATATA	CARG1ATAP3	9.52E-07		S	Binding site of AP3/PI heterodimer	318
r5d	2100	GGGCGGGTGC	AGCBOXNPGLB	1.41E-05		M	Binding sequence of the stress signal-response factors ERFs	4
r5d	2100	G[AG]G[CG]CTCC[CA]A	IDRSZMFER1	4.41E-05		M,S	Iron-Dependent Regulatory sequence	20
r5d	2100	AATAATAT	AT1BOX	1.26E-11		S	Light-regulation of gene expression	229
r2u	1100	[GT]CCACGTG[TG]C	ABRETAEM	0.00E+00		M,S	ABRE (ABA responsive element) found in angiosperms	36
r2u	1100	GGACGGCC[AC][GA]	GRAZMRAB17	7.78E-10		M	Found in the promoter of ABA responsive genes in maize	6
r2u	1100	AAAA[ATC]AAAAA	MARTBOX	8.18E-12		M,S	conserved drought responsive elements in gymnosperm and angiosperms	103
r2u	2100	[CT][TC]GACGGCC[AG]	HEXAMERATH4	1.27E-05		M,S	Found in promoters of plant histone gene H4	9
r2u	2100	CCCAATTGAC	WBOXGACAD1A	1.27E-06		M	WRKY transcription factors binding site	13
r2u	2100	TCGCG[CG]ATCC	OCTAMERMOTIFTAH3H4	4.82E-09		M	Found in promoters of angiosperm histone genes H3 and H4	5
r2u	2100	CGGCCA[AG]ATC	E2FCONSENSUS	1.98E-05		M	E2F-DP-binding motif	11
r2u	1100	CGGATAAG	IBOX	8.49E-06		S	Conserved sequence upstream of light-regulated genes	288
r2u	2100	ATCGTTCG	L4DCPAL1	1.95E-07		S	UV-B responsive element	189
r5u	1100	ATACGCGT	SPHCOREZMC1	1.86E-05		S	maturation program in seed development in maize	462
r5u	2100	CGCGCCGC	ABRECE3ZMRAB28	1.04E-03		S	ABA responsive element	279
r5u	2100	CGCGTGTA	MYCATERD1	9.40E-07		S	Necessary for expression of erd1in dehydrated Arabidopsis	279
r5u	2100	CGCGTGTA	SPHCOREZMC1	9.06E-06		S	maturation program in seed development in maize	279
r5u	2100	CGCGTGTA	NRRBNEXTA	1.28E-04		S	negative regulatory region in Brassica	279

TFBSs with Description highlighted in red are known to regulate drought response genes.

3.2.2. GO terms enrichment results of TFBSs associated with DRTs

We searched for possible association of DRTs sharing specific TFBSs with Gene Ontology biological processes (BPs) terms and KEGG metabolic pathways using annotation data available through the STRING platform (SZKLARCZYK *et al.* 2019). We found that the most (17/24) TFBSs had significant enrichment with BPs GO terms, KEGG pathways, or both (**Table 3.4**). These TFSBs contained significantly more DRTs than the seven TFSBs without enrichment (243.9 vs. 11.7 DRTs; Mann Whitney test, *p*-value=0.00029), indicating a much higher statistical power in enrichment analyses.

Table 3.4 GO Terms and KEGG Pathway enrichment from STRING data

STAMP ID	GO terms Biological Processes	KEGG Pathways	Number of DRTs
PROXBBNNAPA	35	6	74
MARTBOX	35	7	198
SBOXATRBCS	0	0	27
AMMORESIIUDCRNIA1	125	11	491
MYCATERD1	70	8	315
RBCSBOX3PS	70	8	314
CARG1ATAP3	70	8	318
AGCBOXNPGLB	0	0	4
IDRSZMFER1	0	0	20
AT1BOX	0	2	229
ABRETAEM	6	7	36
GRAZMRAB17	0	0	6
MARTBOX	4	0	103
HEXAMERATH4	0	0	9
WBOXGACAD1A	1	1	13
OCTAMERMOTIFTAH3H4	0	0	5
E2FCONSENSUS	0	0	11
IBOX	81	9	288
L4DCPAL1	68	6	189
SPHCOREZMC1	43	4	462
ABRECE3ZMRAB28	27	0	279
MYCATERD1	27	0	279
SPHCOREZMC1	27	0	279
NRRBNEXTA	27	0	279

A total of 1,046 unique DRTs linked to 16 motifs were associated to 213 overrepresented non-redundant BP GO terms (**Table 3.4**). Many of these terms are related to processes known to be involved in drought tolerance and form clusters of functional categories in a REVIGO summary (SUPEK *et al.* 2011), including stress response, root morphogenesis, ion transport, cell wall organization, polysaccharides metabolism and synthesis of secondary metabolites (**Figure 3.1**). Similarly, the top 25 enriched GO terms ranked accordingly to their frequency among the 24 TFBSs contained categories largely associated with response to stress (**Table 3.5**), with 'response to oxidative stress', 'response to stimulus', 'response to stimulus' and 'response to external stimulus' shared by twelve TFBSs.

Table 3.5 Top 25 GO terms by frequency in TFBSs

Description	GO terms	#TFBSs
response to oxidative stress	GO:0006979	13
response to stimulus	GO:0050896	13
response to stress	GO:0006950	12
metabolic process	GO:0008152	12
response to external stimulus	GO:0009605	12
response to chemical	GO:0042221	12
drug catabolic process	GO:0042737	12
oxidation-reduction process	GO:0055114	12
response to oxygen-containing compound	GO:1901700	12
response to acid chemical	GO:0001101	11
response to drug	GO:0042493	11
phenylpropanoid metabolic process	GO:0009698	10
secondary metabolic process	GO:0019748	10
organic substance metabolic process	GO:0071704	10
response to wounding	GO:0009611	9
cellular process	GO:0009987	9
response to bacterium	GO:0009617	8
response to inorganic substance	GO:0010035	8
drug metabolic process	GO:0017144	7
hydrogen peroxide metabolic process	GO:0042743	7
hydrogen peroxide catabolic process	GO:0042744	7
response to antibiotic	GO:0046677	7
multi-organism process	GO:0051704	7
response to other organism	GO:0051707	7
detoxification	GO:0098754	7

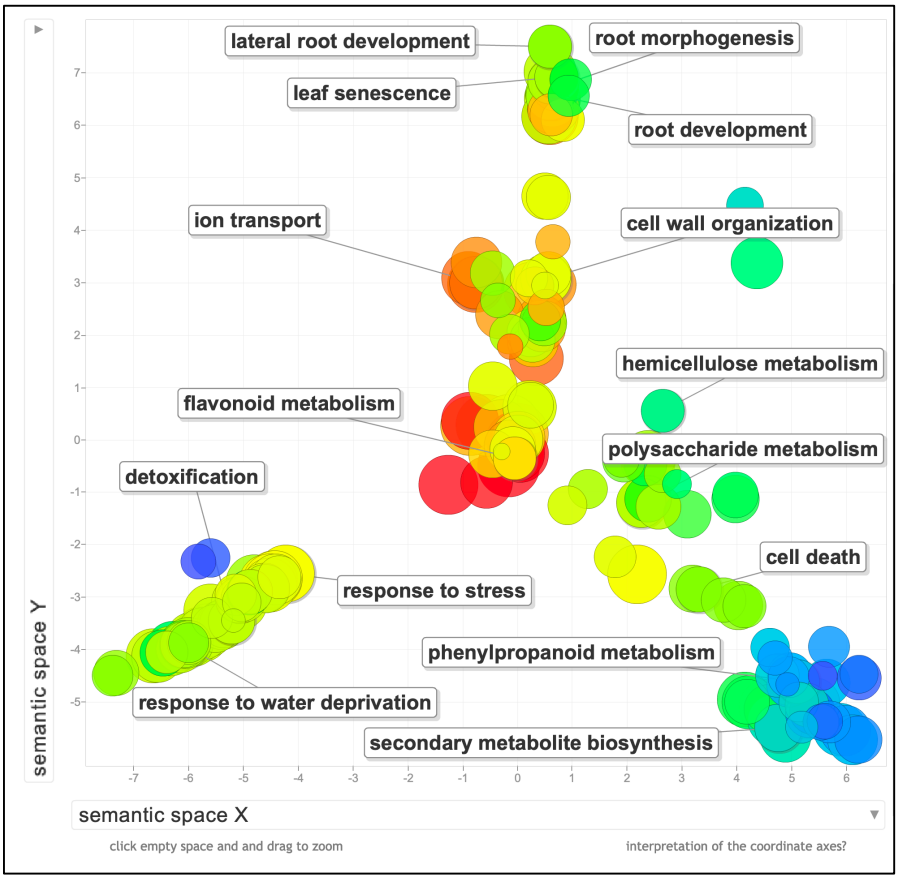


Figure 3.1 REVIGO summary of 213 BP GO terms enrichment for DRTS linked to TFBSs. Highlighted terms are commonly associated with response to aridity, particularly in root.

A total of 147 unique DRTs linked to 12 motifs were found in 24 KEGG metabolic pathways (**Table 3.6**). DRTs from clone 2 were associated to 22/24 pathways, compared to only five pathways associated to clone 5 DRTs. In clone 2, thirteen and twelve pathways were associated with down- and upregulated DRTs, respectively. Most pathways were associated only with one dataset, mirroring the limited functional overlap

of DRTs between genotypes and conditions previously observed in the analysis of all DRTs (Li et al., *unpublished*).

Table 3.6 Summary of KEGG Pathways

KEGG Pathway	#TFBSs	#DRTs	Dataset	Unique DRTs
alpha-Linolenic acid metabolism	4	13	r2d	4
Amino sugar and nucleotide sugar metabolism	3	11	r2d,r2u	8
Ascorbate and aldarate metabolism	4	11	r2d	3
Biosynthesis of secondary metabolites	9	199	r2d,r2u,r5u	79
Carbon metabolism	3	18	r2u	8
Fatty acid degradation	1	4	r2d	4
Flavonoid biosynthesis	5	21	r2d	8
Fructose and mannose metabolism	3	9	r2u	4
Galactose metabolism	1	3	r2u	3
Glutathione metabolism	1	6	r5u	6
Glycolysis / Gluconeogenesis	2	6	r2u	4
Glycosphingolipid biosynthesis	5	10	r2d	2
Glyoxylate and dicarboxylate metabolism	2	7	r2u	4
Linoleic acid metabolism	6	12	r2d	3
MAPK signaling pathway - plant	3	10	r2u	6
Metabolic pathways	9	272	r2d,r2u,r5u	113
Pentose and glucuronate interconversions	1	6	r2d	6
Phenylpropanoid biosynthesis	8	79	r2d,r5d,r5u	31
Plant hormone signal transduction	1	2	r2u	2
Plant-pathogen interaction	1	5	r2d	5
Starch and sucrose metabolism	1	2	r2u	2
Tryptophan metabolism	2	6	r2u	3
Ubiquinone and other terpenoid-quinone biosynthesis	1	3	r5d	3
Zeatin biosynthesis	1	3	r2d	3

Of the four pathways with DRTs from multiple datasets, two correspond to the broad ‘Biosynthesis of secondary metabolites’ and ‘Metabolic pathways’ networks

containing 79 and 113 DRTs linked to nine TFBSs, respectively. Conversely, the two other pathways, 'Amino sugar and nucleotide sugar metabolism' and 'Phenylpropanoid biosynthesis', represent specific metabolic processes. The first pathway has been found in association with drought-induced gene expression changes in several plants (QIU *et al.* 2011; YOU *et al.* 2019). Phenylpropanoids represent secondary metabolites that increase tolerance to mechanical damage and environmental stress, including drought, in both gymnosperms and angiosperms (VOGT 2010). Because these two pathways include up- and downregulated DRTs from the two clones, we sought to determine to what degree the same genes were involved in the two genotypes. As expected, we found no overlap between up- and downregulated DRTs in clone 2 associated with "Amino sugar and nucleotide sugar metabolism". The two groups of clone 2 downregulated DRTs linked to the TFBSs PROXBBNNAPA and AMMORESIIUDCRNIA1, both associated with "Amino sugar and nucleotide sugar metabolism", were entirely overlapping, but differed from the two upregulated DRTs linked to the TFBS ABRETAEM. However, both one upregulated DRT and one downregulated DRT from clone 2 associated with this pathway showed sequence homology with the *A. thaliana* gene *EP3*.

Similarly, upregulated and downregulated datasets associated with "Phenylpropanoid biosynthesis" contained different sets of DRTs, but several of these DRTs shared homology with the same *A. thaliana* genes (**Table 3.7; Figure 3.2**).

Table 3.7 DRTs associated with Phenylpropanoid Biosynthesis

	r2d	r5d	r5u
r2d	17 (16)	2 (4)	0 (3)
r5d	2 (4)	6 (6)	0 (2)
r5u	0 (3)	0 (2)	9 (9)

Numbers in parenthesis represent homologous *A. thaliana* genes

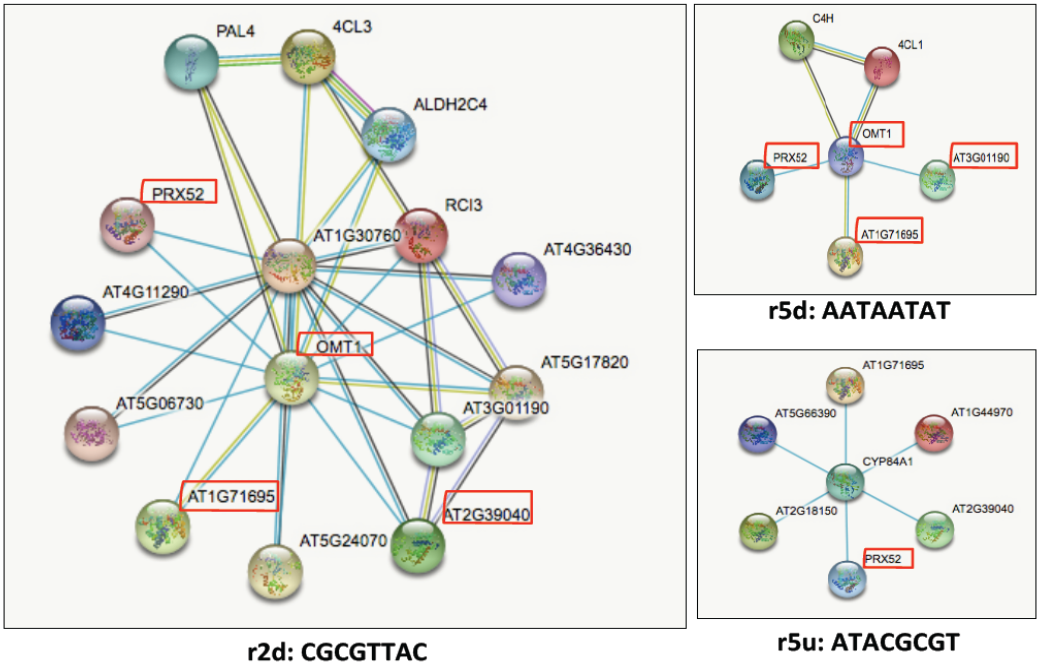


Figure 3.2 Networks of *A. thaliana* genes associated with phenylpropanoid biosynthesis and homologous to DRTs from the datasets r2d, r5d and r5u. Shared genes between clones/expression regimes are highlighted.

3.3. Discussions

Elucidating the origin and function of promoter region transcription factor binding sites (TFBSs) is essential to determine differences in gene expression that lead to phenotypic variation between populations and species. In many nonmodel organisms, large-scale experimental analyses of TFBSs have not been performed. Computational

approaches that leverage on well-studied TFBSs from model species provide the main source of information on the function of putative promoter motifs in nonmodel species. Because TFBSs tend to be formed by DNA sequences that typically encompass only 5-15 nucleotides, it is unclear if TFBSs found in distantly related species share enough sequence similarity to be considered functionally homologous. The computational identification and functional prediction of TFBSs are especially important in plant lineages that are less amenable to functional analyses due to long generation times and limited genetic resources, such as the gymnosperms. Gymnosperms are ecologically prominent in most boreal forests, play an essential role in the forest industry worldwide, and present unique combinations of evolutionary traits. With the recent sequencing and assembly of several gymnosperm genomes, comprehensive bioinformatic analyses of TFBS have become possible in this major group of seed plants.

In this study, we conducted the first large-scale *de novo* computational prediction of TFBSs in gymnosperms focusing on one of the most well studied and economically important gymnosperm, the conifer loblolly pine (*Pinus taeda*, L.). We analyzed thousands of drought related transcripts we have recently discovered using transcriptomic approaches (Li et al., *unpublished*) and identified 24 non-redundant TFBSs in loblolly's putative promoter regions of drought-related genes, the largest number of *cis*-regulatory motifs detected in a single study of gymnosperms thus far. Notably, all the 24 TFBSs correspond to known motifs in angiosperms, highlighting a remarkably widespread conservation of regulatory motifs across the two lineages of seed plants. While this is in line with works showing similarities between a few TFBSs in

gymnosperms and angiosperms (SILVA *et al.* 2015; PEVIANI *et al.* 2016; CHEN *et al.* 2017), our findings revealed that this high level of CRE conservation might be ubiquitous. The lack of gymnosperm-specific motifs suggests that novel TFBSs could have rarely emerged in this lineage. Intriguingly, this also indicates that many angiosperm TFBSs were present in the common ancestor of all seed plants. These results echo the observation of a strong sequence similarity in the ABRE motif between mosses and angiosperms (TIMMERHAUS *et al.* 2011) and warrant more extensive comparisons between the TFBS repertoires of flowering plants and gymnosperms.

One possible explanation for the observed conservation of TFBSs is that drought-related genes might be evolving more slowly than other TFBSs and thus shows higher levels of sequence similarity among seed plants. However, we found that only eleven of the 24 loblolly TFBSs are known to be involved in abiotic stress responses, including response to water deficit, whereas most TFBSs were implicated in other processes. Therefore, it is unlikely that the conservation of TFBSs represents a mere byproduct of the function of the genes analyzed here.

Our results also show that different bioinformatic programs detect largely non-overlapping sets of TFBSs, with only four motifs shared between MEME and Seeder. These discrepancies are most likely due to the different types of algorithms employed by the two programs. MEME is a "sequences driven" method that discovers motifs by calculating the score of position-dependent letter-probability matrix between training set and our motif to find successive motifs (BAILEY AND ELKAN 1994). Conversely, Seeder is a "pattern driven" tool that applies a discerning seeding motif discovery algorithm,

which computes the Hamming distance (HD) of the query sequence and the most associated sequences from a background file to classify the sequences as features to discover motifs (KEICH AND PEVZNER 2002; FAUTEUX *et al.* 2008).

Many of the identified TFBSs are known to be associated with stress response and, in some cases, have been identified in the promoter region of genes involved in drought tolerance (**Table 3.2**). MARTBOX represents a common regulatory motif in the promoter region of monocot and dicot genes (CSERHATI 2015). For instance, it has been found in the promoter regions of drought stress-related gene *JcNAC* in the leaf of physic nut *Jatropha curcas* (WU *et al.* 2015) and played a role of scaffold attachment region in NAC4 transcription factor promoter responsive to environmental stress in *Gossypium hirsutum* (VIKAS SHALIBHADRA TRISHLA 2019). This TFBS is also conserved in the promoter region of the abiotic stress-associated gene *Hsp70* in moss (TANG *et al.* 2016).

Another important motif is ABRETAEM, which was found in promoters of the SK_n, Y_nSK_n and Y_nK_n classes of dehydrins in seed plants, although its possible association with the promoter of gymnosperms' genes has not been determined (ZOLOTAROV AND STROMVIK 2015). ABRETAEM represented the ABRE (ABA responsive *cis*-regulatory element) in wheat (*Triticum aestivum*) (GUILTINAN *et al.* 1990; BUSK AND PAGES 1998). The motif PROXBBNNAPA was found to be involved in seed specific expression and ABA responsiveness in napin *napA* promoter in *Brassica napus* (EZCURRA *et al.* 1999). The TFBS SBOXATRBCS is important for the sugar and ABA responsiveness in Photosynthesis-associated nuclear genes (*PhANGs*) in *A.thaliana* (ACEVEDO-HERNANDEZ *et al.* 2005). MYCATERD1 is a motif that was found to be

involved in the expression of early responsive to dehydration (erd1) in dehydrated *Arabidopsis* (SIMPSON *et al.* 2003; TRAN *et al.* 2004). The AGCBOXNPGLB TFBS acts as binding sequence of the stress signal-response factors ERFs ethylene-responsive element binding factors in *Arabidopsis* (FUJIMOTO *et al.* 2000). GRAZMRAB17 was discovered in the promoter of ABA responsive rab17 gene from maize (*Zea mays*) (BUSK *et al.* 1997). Finally, the motif ABRECE3ZMRAB28 acted as ABA responsive element; stress response in the promoter of rab28 gene in maize (*Zea mays*) (BUSK AND PAGES 1997; BUSK AND PAGES 1998). All of these findings showed conservation of stress related cis-regulatory element in gymnosperm and angiosperms.

AT1BOX is a motif found to regulate the light responsive genes (TERZAGHI AND CASHMORE 1995), and to colocalize with up-regulated genes in drought stress in *Arabidopsis* (HARB *et al.* 2010). The motif S2FSORPL21 was also associated with DRTs showing enriched GO terms. This motif is known to be involved in the expression regulation of genes belonging to the AP2/ERF (APETALA2/Ethylene Responsive Factor) superfamily of transcription factors, which are involved in the response to dehydration and low temperatures (MIZOI *et al.* 2012; CUI *et al.* 2016). S2FSORPL21 was found upstream of the gene *RPL21*, which plays an important role in plastid development and embryogenesis in *A. thaliana* (LAGRANGE T 1997; YIN *et al.* 2012).

The analysis of loblolly transcripts linked with the 24 TFBSs revealed significant functional associations with environmental stress responses, as expected for drought-related genes. Enriched GO term biological processes included response to stress and external stimulus, as well as anatomical and physiological changes in the root system,

the tissue where the gene expression patterns in response to drought were analyzed (Li et al., *unpublished*). KEGG pathways associated with TFBSs included a variety of processes involved in drought tolerance. Multiple secondary metabolites biosynthesis pathways were associated with either up- or downregulated DRTs, in in agreement with previous findings of either increase (BLANCH *et al.* 2009) or decrease (MCKIERNAN *et al.* 2014) of secondary metabolites production between different plants. For example, the 'flavonoids and ascorbate' biosynthesis pathways were both enriched in downregulated DRTs, despite evidence of the increased activity of these two antioxidant pathways during aridity in other plants (NAKABAYASHI *et al.* 2014; AKRAM *et al.* 2017). Conversely, a third antioxidant pathway involved in drought tolerance, 'glutathione metabolism' (NOCTOR *et al.* 2014), was associated to upregulated DRTs. Although these three pathways have been recognized for their roles in drought response in both conifers (FOX *et al.* 2018a) and angiosperm tree species (LEI *et al.* 2006), our results revealed a more nuanced expression regulation of genes involved in antioxidant production.

Even more strikingly, the two pathways 'Amino sugar and nucleotide sugar metabolism' and 'Phenylpropanoid biosynthesis' contained both up- and downregulated DRTs, some of which showed homology with the same *A. thaliana* genes. This finding may be explained by either the evolution, in loblolly, of multiple paralogs of *A. thaliana* single-copy genes associated with aridity tolerance, or the loss of some copies of these genes in *A. thaliana*. Given the rapid gene turnover observed in conifers (CASOLA AND KORALEWSKI 2018), the first scenario appears more probable.

We also observed that several components of carbohydrates metabolism ('Fructose and mannose metabolism', 'Galactose metabolism', 'Glycolysis/Gluconeogenesis', 'Glyoxylate and dicarboxylate metabolism') were associated with upregulated DRTs, as expected for processes that lead to the accumulation of osmoprotectants in response to aridity (LORENZ *et al.* 2011; SINGH *et al.* 2015; MORAN *et al.* 2017).

We argue that the complex relationship between up- and downregulated DRTs, genotypes and metabolic pathways can be explained in light of four arguments. First, promoter regions could not be retrieved from a significant number of DRTs, due to the incomplete annotation of loblolly pine's genes and transcripts. Including more genes in the functional analyses could clarify the association between groups of up- and downregulated DRTs with specific metabolic networks. Second, different combinations of TFBSs can determine different gene expression patterns. Because we analyzed DRTs linked to each TFBS separately, this effect was not accounted for in our results. Third, predictions of functional networks based on distantly related species is potentially confounding the association between groups of up- and downregulated DRTs in nonmodel species and pathways described in model species, as enlightened by the 'Amino sugar and nucleotide sugar metabolism' and 'Phenylpropanoid biosynthesis' cases. Fourth, opposite changes in expression levels between some genes of a metabolic network to the same stimulus are inherent to the biological complexity of the cellular regulatory system, and such nuanced responses might be lost in more coarse clustering analyses.

3.4. Methods

3.4.1. Promoter sequences

Promoter sequences were obtained using the putative transcription start site (TSS) of transcripts characterized as drought-related in our previous analyses of transcriptomic data from loblolly pine seedlings' roots under water stress (Li et al., *unpublished*). Briefly, differentially expressed transcripts from a drought-simulation experiment comparing whole transcriptomes of control and water-deprived loblolly pine (*Pinus taeda* L.) ramets of two clones were used as target sets for the identification of *cis*-regulatory elements. Transcripts either up- or downregulated in water-deprived roots were defined drought-related transcripts, or DRTs. To extract the promoter sequences and identify TFBSs, we used four groups of DRTs: root clone2 upregulated (r2u), root clone2 downregulated (r2d), root clone5 upregulated (r5u) and root clone5 downregulated (r5d). We assumed that in DRTs with an annotated 5' UTR, the first base represented the TSS, and we selected these genes for all subsequent analyses. Because the actual TSS might occur slightly upstream in some transcripts, we retrieved DNA sequences corresponding to 1 kb and 2 kb upstream of the putative TSS from the loblolly pine genome assembly version 1.01 using BEDTools v2.27.1 (QUINLAN AND HALL 2010). Sequences shorter than 100 nucleotides were filtered out. Putative promoter regions with candidate "proximal" (-1,000,+100bp) and "distal" (- 2000,+100bp) sequences upstream of the putative TSS from these DRTs were retrieved in each dataset. Overlapped promoter regions under the same DRTs ID were removed.

3.4.2. De novo identification of TFBSs

To detect TFBSs in the putative promoter regions we searched for motifs occurring at high frequency in each of the four DRT datasets (r2u, r2d, r5u, r5d) using the position-dependent letter-probability matrices approach of MEME (BAILEY *et al.* 2009) and the feature classification implemented in Seeder (FAUTEUX *et al.* 2008). In MEME, background files of each dataset (r2u, r2d, r5u, r5d) were obtained through the program fasta-get-markov, which estimates a Markov model from a control fasta file. The control fasta file was retrieved using two steps. First, a filtered fasta file was obtained by removing the stretches of 'Ns' (gaps) on combined strands from the total fasta files of the promoter regions. Second, the test fasta sequences were deleted to generate the total fasta file. These steps were applied to both promoter regions of 1,100bp and 2,100bp. Each background file with orders from 0 to 3 was used as an input file in the MEME suite. In MEME, such orders represent Markov orders of *k-1,* that is *k*-mer frequencies of a background model file (BAILEY *et al.* 2009). The top 10 motifs were extracted from the background file and each filtered promoter sequence using MEME. Only motifs with e-value≤0.05 were retained. In Seeder, the indexed files with index 6 and index 8 were retrieved using a perl script, then the control background file was obtained through the index file and control fasta file. Finally, the seeder finder was applied for the motif discovery based on the three files including index file, control background file and the test fasta file. Motifs with Q-value≤0.05 were considered significant. TFBS with overlapping genomic coordinates were identified and redundant motifs with any overlap (1bp or more) were removed.

3.4.3. Prediction of TFBSs function

Conservation of the discovered TFBSs among seed plants was investigated using the STAMP (MAHONY AND BENOS 2007) web tool. The PLACE database (HIGO *et al.* 1999) was used for similarity analyses and motif functional annotation. E-value thresholds of 0.005 and 0.002 were used in STAMP similarity results for MEME motifs and Seeder motifs, respectively. The functional annotations were then searched and added to the retained motifs. The STAMP results were used to remove redundant TFBSs and to identify TFBSs annotated by both MEME and Seeder.

3.4.4. Functional enrichment analysis of TFBS related genes

Searches on the *A. thaliana* STRING database [22] were performed for all transcripts associated with each significant TFBS using protein sequences. Enriched biological processes GO terms and KEGG Pathways at FDR≤0.05 were used in functional analyses.

3.4.5. Data

Data related to this project including DNA sequences of promoter regions, protein sequences have been deposited on the Figshare repository.

4. ADAPTATION IN LOBLOLLY PINE DROUGHT-RELATED GENES

4.1. Introduction

Local adaptation is associated with allele frequency changes between populations (GUNTHER AND COOP 2013). Genes involved in the response to abiotic phenomena, including water availability, may thus expected to be more likely to experience shifts in allele frequencies. These shifts can be identified through statistical tests, such as heterozygosity testing, *F* statistics, Nei's genetic distance, population assignment, probabilities of identity and pairwise relatedness, which can identify levels of deviations from expectation of neutrality in polymorphic markers. The widespread access to high-throughput DNA sequencing resources in the past two decades has enabled correlation studies of local adaptation in humans and other species, including forest trees (HANCOCK *et al.* 2010; PRITCHARD *et al.* 2010; NEALE AND KREMER 2011; LE CORRE AND KREMER 2012).

Common analyses to detect genotype-phenotype associations are represented by tests that identify departures of nucleotide variation patterns from the expectation of the molecular theory of neutral evolution. These tests include Tajima's *D*-statistic, which compares the estimates of the number of segregating sites and the mean pairwise difference between sequences (TAJIMA 1989). Other commonly used statistics include the Fu and Li's *F** (FLF*) based on the number of derived singleton mutations and the mean pairwise difference between sequences (FU AND LI 1993) and Fu and Li's *D** (FLD*), which compares the number of derived singleton mutations and the total

number of derived alleles (FU AND LI 1993). Other tests rely on comparison between allele frequency within a species with estimates of nucleotide divergence with an outgroup (sister) species. For instance, the Fay and Wu's *H*-test compares the relative excess of low- and high-frequency-derived alleles with the number of variants immediately after a selective sweep (FAY AND WU 2000), and the Hudson-Kreitman-Aguade (HKA) test (HUDSON *et al.* 1987), which compares the polymorphism within species and the divergence between species in a particular region.

F_{ST} is a widely used statistic applied to population genetic data to estimate changes in allele frequency between populations. F_{ST} represents the genetic variance in a subpopulation relative to the total population and is often used to discern signatures of balancing or positive selection in SNPs (FLANAGAN AND JONES 2017). Though there are potential false positive issues related to the use of the F_{ST} outliers detection method (WHITLOCK AND LOTTERHOS 2015), the outlier SNPs showing significant association with environmental variables form a reliable set of variants implicated in local adaptation (LU *et al.* 2019).

Given the key role played by water availability in plant life history, understanding the genetic basis of adaptation to aridity has received much attention in plant biology (SEKI *et al.* 2001; MCKAY *et al.* 2003; PELEG *et al.* 2008; HADIARTO AND TRAN 2011; STEANE *et al.* 2014; STEANE *et al.* 2015; YE *et al.* 2017; BARTON *et al.* 2020). For example, STEANE *et al.* 2014 leveraged Bayesian analysis to identify a set of 94 putatively adaptive sequence-tagged markers across the genome of *Eucalyptus*

tricarpa trees from 9 provenances in southeastern Australia that were strongly correlated to temperature and water availability at their original sites.

In conifers, a number of studies have been conducted on signature of adaptation with tolerance to aridity (GONZALEZ-MARTINEZ *et al.* 2008; ECKERT *et al.* 2010). Using several different tests including nucleotide diversity, Tajima's *D*-test, Fu's *Fs*-test, Fay and Wu's *H*-test statistics, GONZALEZ-MARTINEZ et al. (2006) identified a SNP associated with aridity tolerance in the gene *erd3* (early-response-to-drought-3), and significantly lower values of haplotype diversity in candidate genes *lp3-3* (water-stress inducible protein 3), *ferritin*, *pp2c* (protein-serine/threonine phosphatase) and *ccoaomt-1* (caffeoyl-CoA O-methyltransferase 1). Notably, variants in the gene *erd3* were also detected in several other studies as associated to drought tolerance in loblolly pine (ERSOZ *et al.* 2010; KORALEWSKI *et al.* 2014). The genes *ug-2_498* and *ppap 12* (putative wall-associated protein kinase) showed significant Tajima's *D* values; *ug-2_498* with negative Tajima's *D* value indicating possible positive selection and gene *ppap 12* with positive Tajima's *D* value indicating possible balancing selection (GONZALEZ-MARTINEZ *et al.* 2006). Eckert and collaborators found five SNPs associated with aridity in the ADEPT2 population using an array of 3,059 SNPs. These SNPs were found in genes encoding an hexose membrane transporter, a photosystem II protein, a C3HC4-type RING finger transcription factor, a MATE transporter, and a UDP-galactose transporter (ECKERT *et al.* 2010). The gene *cad* (carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase) showed evidence of positive selection or balancing selection by significant positive Tajima's *D* and *agp-6*

(arabinogalactan protein 6) was found to evolve under positive selection with a significant negative Tajima's *D* (KORALEWSKI *et al.* 2014).

There are also SNPs found in other conifers related to adaptation to drought. Variants in the two genes *PR-AGP4* and *erd3* have been found to be associated with diversifying selection in maritime pine (*Pinus pinaster* Ait.) (EVENO *et al.* 2008). In Scots pine (*Pinus sylvestris*) and maritime pine, the genes *dhn2*, *dhn5* and *col1* were identified as involved in adaptation to abiotic stress (GRIVET *et al.* 2017). Moreover, seven SNPs were detected in loci associated with environmental variables including aridity indices, precipitation and mean diurnal range temperature in the Aleppo pine (*Pinus halepensis*) (RUIZ DANIELS *et al.* 2018). These SNPs were identified within the genes encoding the Peroxisomal membrane protein 11D-like, Polygalacturonase inhibitor 1-like, Alpha-galactosidase-like, RING-H2 finger protein ATL48-like, B3 domain-containing protein At3 g19184-like and with one unknown protein (RUIZ DANIELS *et al.* 2018).

Genome-wide analyses of genetic variants have recently become available in loblolly pine and a few more conifers. For instance, large-scale datasets of polymorphisms have recently become available in loblolly via exome-based genotyping analyses (LU *et al.* 2016), enabling the identification of a high number of polymorphisms associated with traits, climate variables or genes known to be involved in drought tolerance. Using 87,000 SNPs obtained from genome resequencing data (DE LA TORRE *et al.* 2018), De La Torre and collaborators also reported that water availability represents the primary climate variable associated with local adaptation in loblolly pine

(DE LA TORRE *et al.* 2019). Several SNPs have been found to be associated with aridity in these and other studies focused on loblolly pine and other Pinaceae, mostly based on F_{ST} outlier approaches.

For instance, (DE LA TORRE *et al.* 2019) applied a hierarchical clustering method to search for significant associations between SNP allele frequencies and environmental variables and identified six SNPs associated with temperature-related variables, with three of them located in the same scaffold in loblolly pine. ECKERT *et al.* (2013) tested nucleotide diversity using amplicons across the loblolly pine genome, and discovered amplicons associated with drought had the lowest nuleotide diversity compared to other categories of amplicons, such as expression levels for lignin and cellulose-related genes, primary metabolite concentrations, and disease resistance (ECKERT *et al.* 2013). LU *et al.* showed abiotic stress responsive genes encoding Asparagine synthetase, 2-oxoisovalerate dehydrogenase, late embryogenesis abundant protein, WAT1-related protein, bark storage protein A-like among 611 environmentally associated SNPs (LU *et al.* 2019).

These studies have enabled the discovery of some genetic variants and genes associated with environmental variables, including aridity, in loblolly pine and other conifers. In agreement with a large body of work on plant adapation to drought, I have shown in Chapter 2 that thousands of loblolly genes significantly change their expression in the response to drought, and possibly most of them are directly implicated in the genetic basis of adaptatioon to aridity in this species. Therefore, a comprehensive understanding of the genetic and molecular basis of drought response and drought

tolerance requires the integration of association approaches and polymoprhism datasets (F_{ST} outliers, large collections of SNPs) with functional methods (i.e. transcriptomic analyses). In this Chapter, I sought to highlight the value of combining population genomic resources and association data with the large-scale analysis of gene differential expression to drought presented in Chapter 2. To this end, I followed two lines of investigation based on two hypotheses. First, I tested the hypothesis that DRTs should contain more SNPs known to be associated to drought compared to non-DRTs. My second hypothesis is that DRTs evolve under stronger positive selection as opposed to non-DRTs, given the high levels of genetic diversity and strong selective pressure due to abiotic factors, especially water availability, in *Pinus taeda*.

4.2. Results

4.2.1. Identification of variants associated with aridity in DRTs and non-DRTs

I investigated the frequency of DRTs and non-DRTs that contain SNPs known to be associated with aridity or other climate variables, and SNPs representing F_{ST} outliers. To map these SNPs to the transcriptome, I performed sequence similarity searches using BLAST (CHAMACHO *et al.* 2010) between transcripts and sequences containing these SNPs (see **Methods**). A total of 58 such SNPs were identified in DRTs as opposed to 378 in non-DRTs, a statistically significant difference (Fisher's exact test, p=0.0252). Significantly more SNPs were also found for outlier SNPs identified by LU *et al.* (2019) using the two programs SPA and OutFLANK, and for the number of transcripts with SNPs identified by OutFLANK and by Samβada (**Table 4.1**). No significant differences

were observed in the number of SNPs or transcripts with SNPs for variants associated with climate (TASSEL) or with metabolite or expression level changes.

Table 4.1 Outlier SNPs and SNPs associated with climate in DRTs and non-DRTs

	SNPs		Transcripts	
	DRTs	non-DRTs	DRTs	non-DRTs
TASSEL SNPs	4	66	4	64
SPA Outliers	49*	314	15	142
OutFLANK Outliers	7*	4	6*	1
Samβada Outliers	9	50	7*	31
Metabolite levels	1	24	1	19
Expression levels	4	38	3	32
Associated SNPs	58*	378	23	204

Asterisks show higher than expected proportions of SNPs or number of transcripts with SNPs in DRTs vs. non-DRTs at p-value<0.05.

4.2.2. Signatures of natural selection in DRTs and non-DRTs

To detect signatures of selection in DRTs and non-DRTs, I first mapped the 2,822,609 SNPs from the loblolly exome capture and sequencing data reported in LU *et al.* (2016) onto coding regions, 3 prime UTRs and 5 prime UTRs of the 60,090 transcripts reported in Chapter 2 using genome coordinates in gff files (**see Methods**). I identified 83,633 SNPs mapping to DRTs and 752,656 SNPs mapping to non-DRTs, of which 82,509 and 539,159 map to the CDS, respectively (**Table 4.2**). These SNPs mapped onto 2,361 and 26,616 DRTs and non-DRTs, respectively (**Table 4.2**). SNPs were found in ~59% of DRTs compared to 47.5% of non-DRTs (**Table 4.2**). DRTs also showed a higher number of SNPs/transcript (after correcting for transcript length, 30.8

vs. 19.2 SNPs/transcripts; **Fig. 1**) and a significant excess of nonsynonymous variants (Fisher's exact test, p=0.0105). This finding suggests that drought-related genes evolved under a stronger positive selection regime or experienced lower levels of natural selection (**Table 4.2**). The distribution of the number of SNPs per transcript indicated that a higher proportion of DRTs with SNPs for a number of SNPs per transcripts between 10 and 20 (**Fig. 4.1**).

Table 4.2 Features of DRTs and non-DRTs with SNPs

	DRTs	**non-DRTs**
Total SNPs (CDS and UTRs)	83,633	752,656
Total SNPs (CDS-only)	82,509	539,159
#Synonymous SNPs	27,502	182,150
#Nonsynonymous SNPs	55,007	357,009
Nonsyn/Syn SNPs	2.00	1.96
#Transcripts	2,361	26,616
#Genes	2,042	17,041
Average #SNPs/Transcript	34.95	20.26
Average #SNPs/Gene	40.41	31.64
Median #SNPs/Transcripts	27	20
Average transcript length	1136.36	1055.85
Total transcript length	2,682,945	28,102,464
SNPs/1000bp	30.75	19.19
#SNPs per transcript	1-390	1-521
Total transcripts	4,012	56,078
Total genes	3,495	31,641
%Transcripts w/SNPs	58.85	47.46
%Genes w/SNPs	58.43	53.86
Transcripts/Gene	1.15	1.77
Total #Transcripts combined	**60,090**	
Total #Genes combined	**35,136**	

Figure 4.1 Distribution of the number of SNPs per transcripts.

To obtain more direct estimates of selection and molecular evolution, I calculated several diversity and allele frequency distribution statistics using the program PopGenome (PFEIFER *et al.* 2014), including Watterson's estimator of genetic diversity, θ_W, Tajima's estimator θ_T, Tajima's *D* and Fu and Li's *D**, and the ratio of nonsynonymous to synonymous polymorphisms (N/S). To avoid estimating these parameters multiple times using the same sets of SNPs shared by alterative transcripts of the same genes, I only included the longest transcript for each locus with multiple transcripts. These led me to estimate the population genetic parameters for a total of 1,544 DRTs and 16,789 non-DRTs (**Table 4.3**).

DRTs showed higher average nucleotide diversity according to both θ_W and θ_T (**Table 4.3**). Accordingly, the distribution of θ_W is skewed toward higher values in DRTs compared to non-DRTs (**Fig. 4.2**). Additionally, DRTs showed higher average Tajima's

D and Fu and Li's *D** (**Table 4.4, Figs. 4.3-4.4**). N/S values were comparable between DRTs and non-DRTs (**Table 4.3, Figs. 4.5**). These differences were also observed when up- and downregulated transcripts were analyzed separately, with more pronounced differences between upregulated DRTs and non-DRTs in N/S (**Table 4.3**). However, downregulated DRTs average θ_W is significantly higher (*p*-value = 0.004) than upregulated DRTs. Tajima's *D* was twice as high as in downregulated non-DRTs, while average Tajima's *D* of upregulated DRTs was 1.5 higher than that of non-DRTs (**Fig. 4.3**). Also, downregulated DRTs showed significantly higher (*P*-value = 0.0008) average Tajima's *D* than upregulated DRTs, and accordingly the distribution of Tajima's *D* values was skewed toward higher values in downregulated DRTs (**Fig. 4.3**).

Table 4.3 Summary of neutrality test among different set of transcripts

	#Transcripts	θ_W	θ_T	#Segregating sites	Tajima's *D*	Fu and Li *F**	Fu and Li *D**	Total SNPs	S	N	N/S
All transcripts	18,333	6.020	6.464	43.321	0.152	1.390	1.997	642661	208542	434119	2.08
DRTs all	1,544	7.169	7.993	51.595	0.273	1.555	2.188	66820	21319	45501	2.13
Non-DRTs all	16,789	5.914	6.323	42.560	0.141	1.375	1.980	575841	187223	388618	2.08
DRTs upregulated	603	6.922	7.425	49.813	0.173	1.482	2.150	24476	7827	16649	2.13
Non-DRTs upregulated	8,274	5.815	6.176	41.845	0.117	1.358	1.968	274257	90638	183619	2.03
DRTs downregulated	941	7.328	8.357	52.738	0.338	1.602	2.213	42344	13492	28852	2.14
Non-DRTs downregulated	8,515	6.010	6.466	43.254	0.165	1.392	1.991	301584	96585	204999	2.12
r2u 662	272	6.191	6.489	44.555	0.156	1.470	2.131	9344	3058	6286	2.06
r5u 1391	377	7.707	8.328	55.467	0.170	1.496	2.196	17545	5549	11996	2.16
r2d 1041	581	7.163	8.240	51.549	0.366	1.629	2.225	25046	8099	16947	2.09
r5d 981	361	7.645	8.777	55.017	0.304	1.588	2.226	17137	5393	11744	2.18

Fu and Li's *D** did not present a significant difference between up- and downregulated DRTs, and the distribution of Fu and Li's *D** is also higher in larger values for DRTs than non-DRTs (**Fig. 4.4**). The N/S distribution was similar on average between up- and downregulated DRTs and non-DRTs (**Fig. 4.5**).

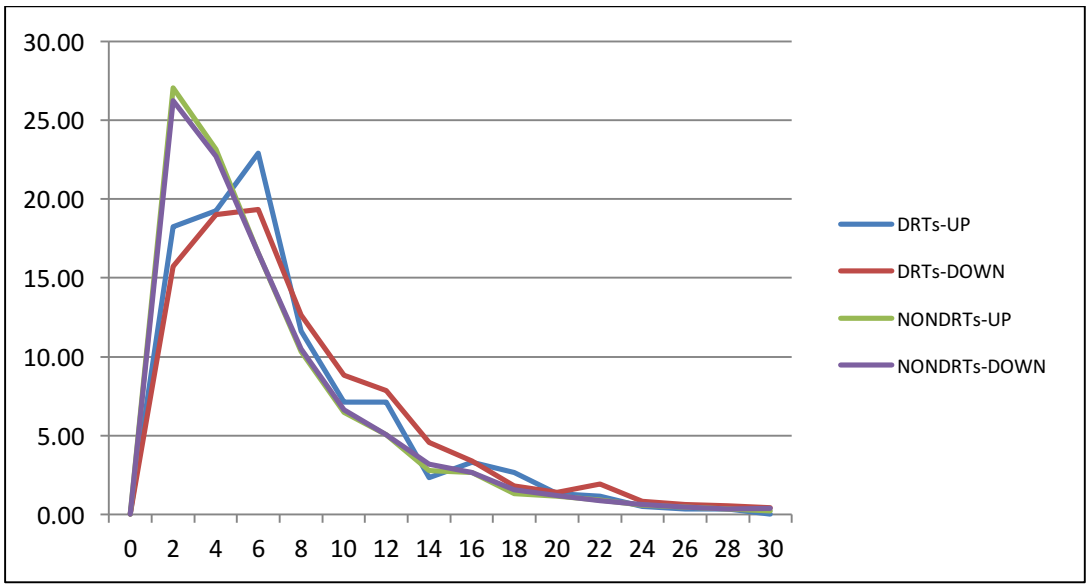


Figure 4.2 Watterson's θ distribution in up- and downregulated DRTs and non-DRTs.

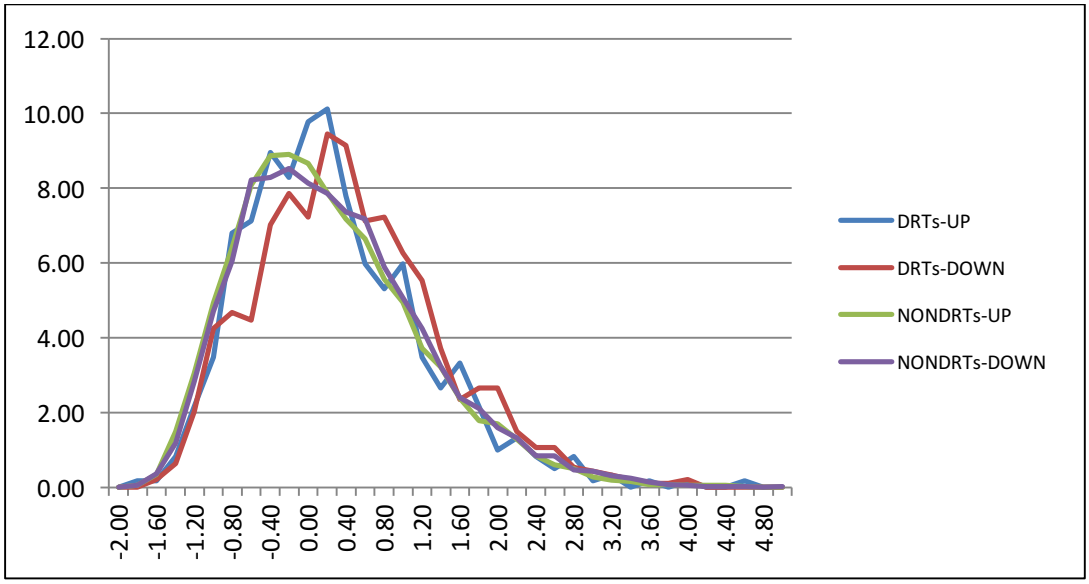


Figure 4.3 Tajima's *D* distribution in up- and downregulated DRTs and non-DRTs.

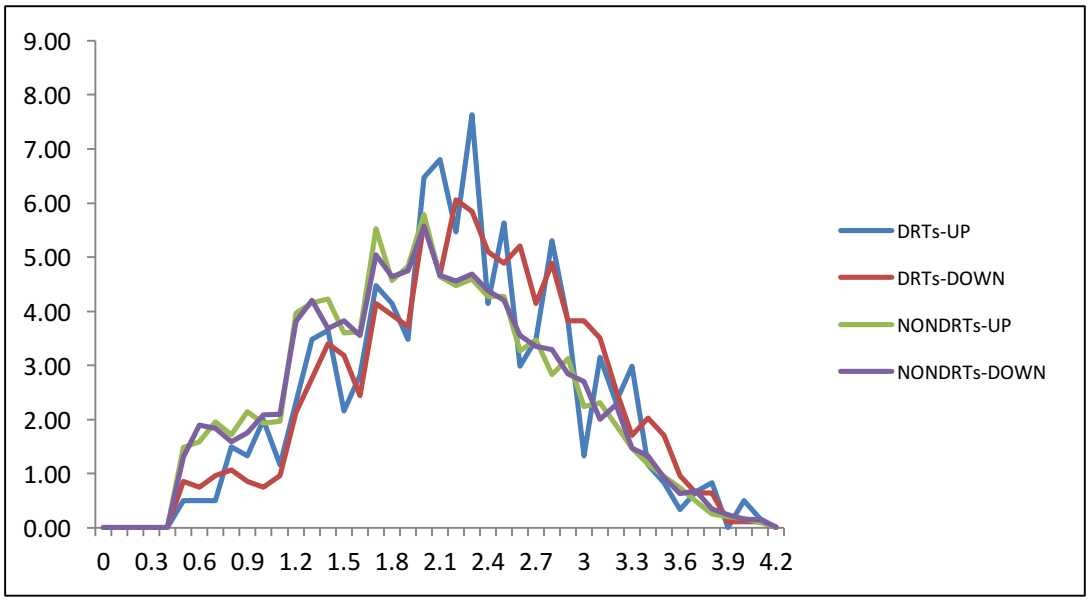


Figure 4.4 Fu and Li's *D** distribution in up- and downregulated DRTs and non-DRTs.

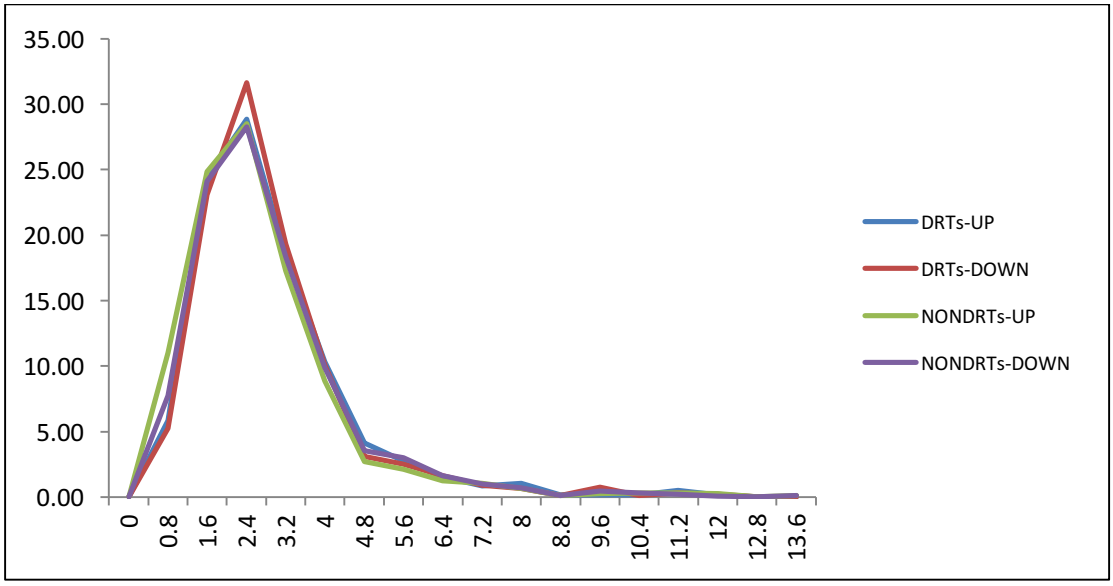


Figure 4.5 N/S distribution in up- and downregulated DRTs and non-DRTs.

I further analyzed possible differences in nucleotide diversity and allele frequency distributions between DRTs of the two loblolly genotypes, clone 2 and clone 5. I found that on average both θ_W and θ_T were significantly higher in clone 5 than clone 2, particularly in upregulated DRTs (**Table 4.3**, **Fig. 4.6**). Indeed, he average θ_W was significantly higher in clone 5 (**Table 4.4**). Tajima's *D* was higher in downregulated DRTs of both clones, particularly in clone 2 (**Table 4.3; Fig. 4.7**). Notably, though both Fu and Li's *D** and N/S described the same trend as θ_W, with DRTs in clone 5 showing higher average values than clone 2, although not in a statistically significant way (**Figs. 4.8-4.9**).

Table 4.4 Neutrality statistics of clone 2 vs. clone 5

	P-value	Average in clone 2	Average in clone 5
θ_W	0.0153	6.8531	7.6767
Fu and Li's *D**	0.6789	2.1950	2.2107
N/S ratio	0.7169	2.4921	2.5261
Tajima's *D*	0.1871	0.2993	0.2351

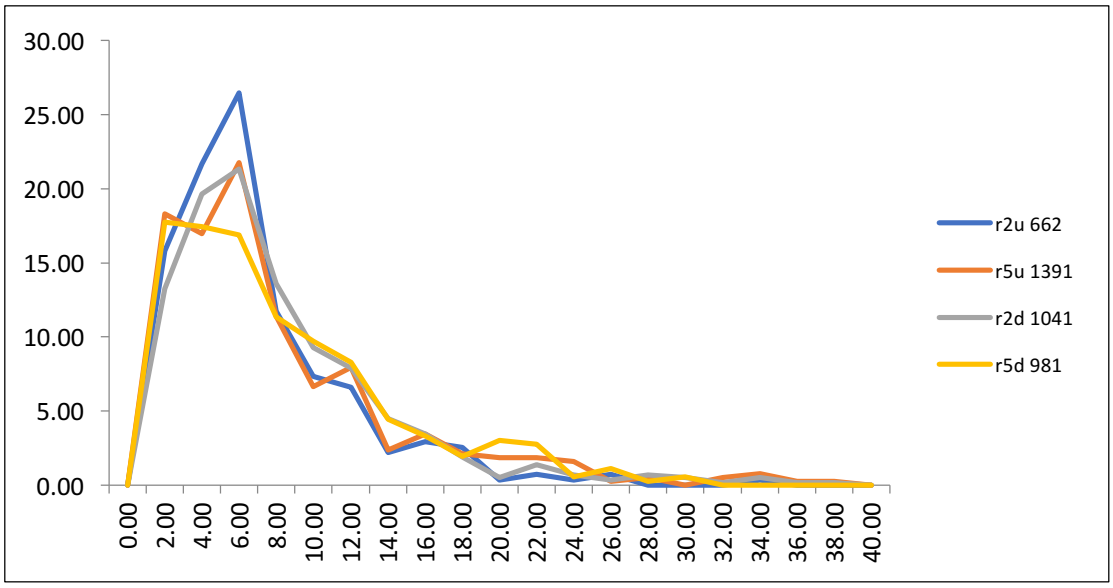


Figure 4.6 Watterson's θ distribution in clone 2 and clone 5 DRTs.

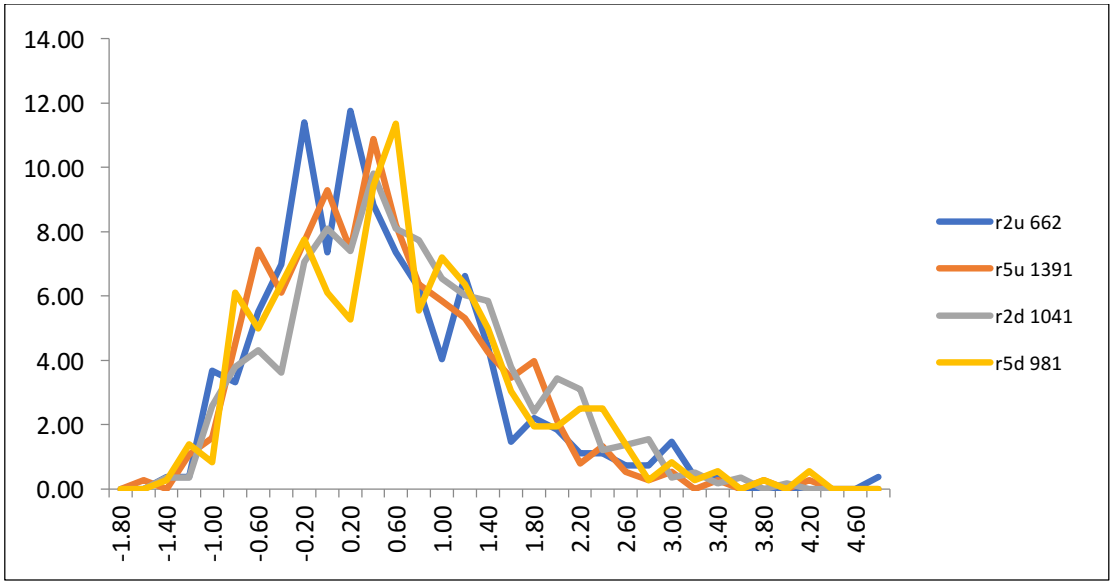


Figure 4.7 Tajima's *D* distribution in clone 2 and clone 5 DRTs.

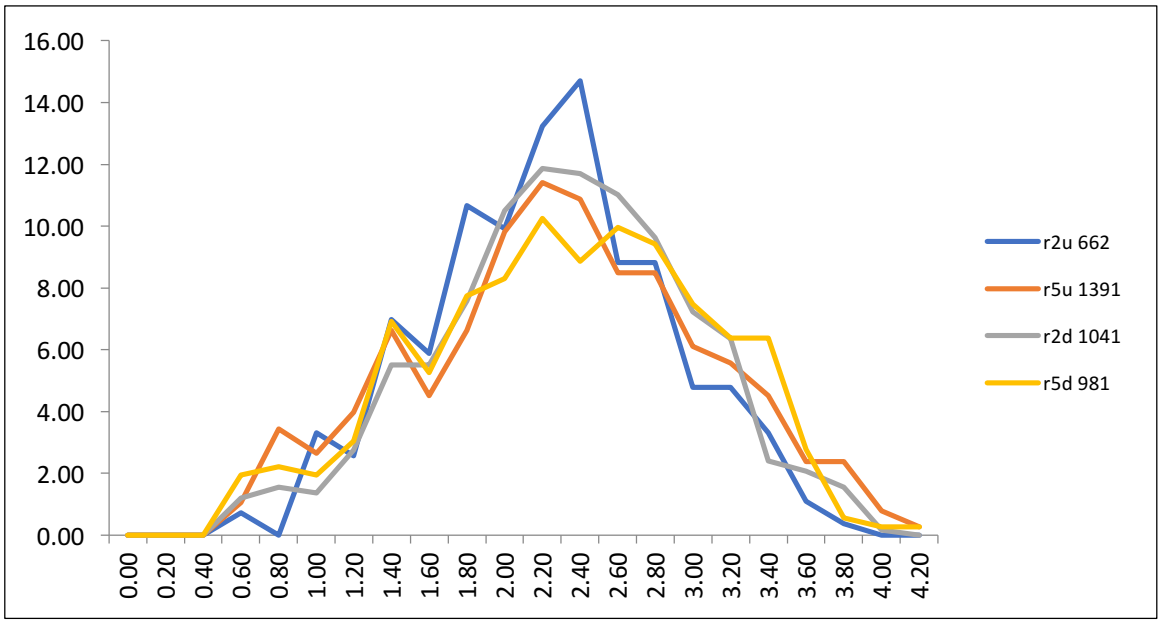


Figure 4.8 Fu and Li's *D** distribution in clone 2 and clone 5 DRTs.

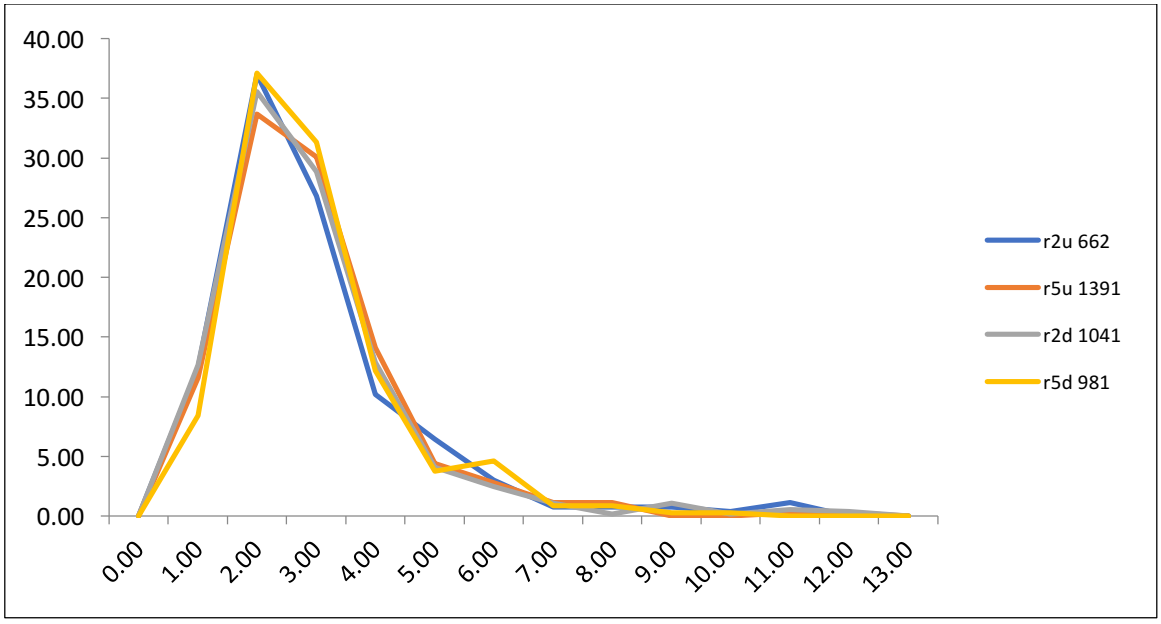


Figure 4.9 N/S distribution in clone 2 and clone 5 DRTs.

4.3. Methods

4.3.1. SNPs association with transcripts and file preparation for PopGenome

A total of 2,822,609 SNPs discovered by Lu et al. were associated by mapping to all the 60,090 transcripts found in chapter 2 (Lu *et al.* 2016; Lu *et al.* 2017). First, the coding region, 3 prime UTR and 5 prime UTR were fetched from the annotation gff3 file of all transcripts in *Loblolly Pine* transcriptome. Second, the complete SNPs vcf file was mapped to the extracted transcript fragments annotation file by python script according to the transcript ID and position. The sequence file for the SNPs associated transcripts was retrieved from the transcriptome fasta file. The corresponding annotation file was obtained from the transcriptome annotation file. Both of the fasta file and gff3 file were extracted using python scripts.

4.3.2. Neutrality test with SNPs associated transcripts

PopGenome (PFEIFER *et al.* 2014) was used to estimate levels of nucleotide diversity and perform neutrality test on coding sequences of transcripts. All three SNP-associated transcripts files were formatted according to the input requirement of the PopGenome. Tajima' *D*, Watterson's estimator of θ, θ_W, the Tajima's estimator of θ, θ_Γ, Fu and Li's F^* (FLF) and Fu and Li's D^* (FLD) values for each SNP-associated transcript were calculated by PopGenome. SNPs associated transcripts were further matched to the up- and downregulated DRTs and non-DRTs, and up- and downregulated clone2 and clone5 varieties from Chapter 2.

4.3.3. Identification of natural selection in DRTs and non-DRTs

Two complementary approaches were conducted to determine the impact of natural selection on DRTs. First, I investigated the frequency of DRTs and non-DRTs that contain SNPs known to be associated with aridity or other climate variables, and SNPs representing F_{ST} outliers. Using a BLASTn analysis comparing transcript sequences with sequences containing these variants, I assigned SNPs to transcripts from the transcriptome described in Chapter 2. In my second approach, I used the SNPs detected by Lu et al. (LU *et al.* 2016; LU *et al.* 2017) and assigned to transcripts to investigate on the evolutionary dynamics of DRTs across 384 loblolly trees from the ADEPT2 population.

4.4. Discussion

Genes involved in local adaptation are expected to evolve towards high levels of nucleotide divergence between populations due to divergent selective pressures. Variation in water availability and temperature are major forces shaping adaptation to climate across a species' range, thus genes involved in physiological responses to changes in these variables might be experiencing high levels of divergence in species whose distribution encompass ecosystems with significant variation in precipitation and seasonal temperatures, such as *Pinus taeda*. In this Chapter, I tested this hypothesis using two complementary approaches. First, I found significantly more SNPs known to be associated with aridity-related environmental variables in DRTs than non-DRTs. This finding is in agreement with previous works on other plants. For instance, a study in the C_4 perennial grass *Panicum hallii* (Poaceae) showed that temperature and aridity associated SNPs were more frequently found in or near genes associated with drought recovery process (GOULD *et al.* 2018).

Second, I determined that on average DRTs experience both higher levels of nucleotide diversity and deviation from neutrality than non-DRTs. Specifically, DRTs showed higher θ, Tajima's *D*, Fu and Li's *D**, and N/S. The increased nucleotide diversity is in line with expectation of more elevated diversity in genes involved in local adaptation between populations. The distribution of Tajima's *D* in DRTs was skewed towards positive values, indicating that the proportion of DRTs evolving under balancing selection is higher compared to non-DRTs. Higher Tajima's *D* have also been found in abiotic stress response genes compared to reference genes in maritime pine

(GRIVET *et al.* 2017). Interestingly, Tajima's *D* was particularly elevated in loblolly pine downregulated DRTs, wherein little difference was found in average Tajima's *D* values between up- and downregulated non-DRTs.

Additionally, up- and downregulated DRTs showed higher average N/S values than upregulated non-DRTs, but comparable values in downregulated non-DRTs. Elevated nonsynonymous single-nucleotide polymorphisms and divergence in the abiotic stress-responsive genes compared to reference genes have also been described in *Solanum chilense* (BONDEL *et al.* 2018).

At the genotype level, I found higher levels of nucleotide diversity in clone 5 vs. clone 2. Clone 5 represents the more drought-tolerant genotypes and might experience increased selective pressure and adaptation at the gene level, as also indicated by the higher N/S values of both up- and downregulated transcripts in this clone. On the contrary, Tajima's *D* values did not differ remarkably between the two clones. However, Tajima's *D* was again higher in downregulated transcripts. This possibly indicates that both clones experience similar levels of balancing selection.

5. CONCLUSIONS

The combination of reduced precipitation and increased temperatures due to climate change and repurposing of water resources is increasingly exacerbating the intensity and duration of drought condition across many ecosystems. This is expected to severely impact the productivity and health of natural and commercial forests. Understanding the genetic and molecular basis of water tolerance in key forest species has the potential to accelerate the development of genotypes with improved drought tolerance that can withstand the projected increase in aridity in many areas.

The goal of this research was to develop an increased knowledge of the genetic basis of drought tolerance as well as genomic resources to promote further work on this topic in loblolly pine (*Pinus taeda* L.), a primary forest species in the southeastern U.S. Previous studies have provided important information about genes involved in drought in loblolly. For example, SPERRY *et al.* 2002 identified a genetic component to the responsiveness of xylem morphology and leaf-level physiology to drought. Studies based on microarray data have revealed some components of the genetic networks involved in drought response of this pine species (LORENZ *et al.* 2011) and other conifers (MORAN *et al.* 2017). Next-generation sequencing technologies, such as RNA-seq, can generate more comprehensive information of genes and transcripts implicated in drought tolerance, but had not been applied to loblolly pine yet. Additionally, the availability of the loblolly pine genome has significantly improved the ability to identify both putative

regulatory regions with a role in drought-related genes expression, and genetic variants associated with environmental variables or traits related to aridity.

Motivated by these facts, I have focused my dissertation research to conduct a systematic study to identify differentially expressed genes involved in drought response, their transcription factor binding sites (TFBSs), and the selection regime acting on these genes. Specifically, I investigated the genetic basis of response and tolerance to low water availability in loblolly pine in three major areas. First, I identified genes and genetic networks associated with drought tolerance by comparing changes in expression patterns between loblolly varieties under simulated drought conditions using RNA-seq data. Second, I discovered novel regulatory regions in the promoter regions of drought-related genes. Third, I assessed the signature of adaptation that correlated with evolution of drought-related genes in loblolly pines. As a result, I showed that remarkably different genetic networks are involved in the response to drought between loblolly varieties. The identification of an array of TFBSs conserved between loblolly and angiosperms implies that *cis*-regulatory motifs are shared between distantly related seed plants. Finally, the hypothesis that drought-related genes are evolving rapidly was validated. These findings will help to better understand the genetic basis underlying drought-resistant of loblolly pine genotypes, assisting in breeding efforts and forest management, and facilitating further studies on the regulatory framework and the role of adaptation in the evolution of stress related genes in plants.

REFERENCES

(IPCC), I. P. o. C. C., 2013 Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. New York: Cambridge University Press 1535 p.

Acevedo-Hernandez, G. J., P. Leon and L. R. Herrera-Estrella, 2005 Sugar and ABA responsiveness of a minimal RBCS light-responsive unit is mediated by direct binding of ABI4. Plant Journal 43**:** 506-519.

Agriculture, U. S. D. o., 2016 Forest Inventory and Analysis.

Aitken, S. N., S. Yeaman, J. A. Holliday, T. Wang and S. Curtis-McLane, 2008 Adaptation, migration or extirpation: climate change outcomes for tree populations. Evol Appl 1**:** 95-111.

Akram, N. A., F. Shafiq and M. Ashraf, 2017 Ascorbic acid-a potential oxidant scavenger and its role in plant development and abiotic stress tolerance. Frontiers in plant science 8**:** 613.

Alter, S., K. C. Bader, M. Spannagl, Y. Wang, E. Bauer *et al.*, 2015 DroughtDB: an expert-curated compilation of plant drought stress genes and their homologs in nine species. Database (Oxford) 2015**:** bav046.

Ambawat, S., P. Sharma, N. R. Yadav and R. C. Yadav, 2013 MYB transcription factor genes as regulators for plant responses: an overview. Physiology and Molecular Biology of Plants 19**:** 307-321.

Andrews, S. F. A. Q. C. T. f. H. T. S. D. O., 2010 FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant *et al.*, 2009 MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37**:** W202-208.

Bailey, T. L., and C. Elkan, 1994 Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2**:** 28-36.

Baldoni, E., P. Bagnaresi, F. Locatelli, M. Mattana and A. Genga, 2016 Comparative Leaf and Root Transcriptomic Analysis of two Rice Japonica Cultivars Reveals Major Differences in the Root Early Response to Osmotic Stress. Rice (N Y) 9**:** 25.

Baldoni, E., A. Genga and E. Cominelli, 2015 Plant MYB Transcription Factors: Their Role in Drought Response Mechanisms. Int J Mol Sci 16**:** 15811-15851.

Barton, K. E., C. Jones, K. F. Edwards, A. B. Shiels and T. Knight, 2020 Local adaptation constrains drought tolerance in a tropical foundation tree. Journal of Ecology 108**:** 1540-1552.

Behringer, D., H. Zimmermann, B. Ziegenhagen and S. Liepelt, 2015 Differential Gene Expression Reveals Candidate Genes for Drought Stress Response in Abies alba (Pinaceae). PLoS One 10**:** e0124564.

Berardini, T. Z., L. Reiser, D. Li, Y. Mezheritsky, R. Muller *et al.*, 2015 The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis 53**:** 474-485.

Birol, I., A. Raymond, S. D. Jackman, S. Pleasance, R. Coope *et al.*, 2013 Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29**:** 1492-1497.

Blanch, J.-S., J. Penuelas, J. Sardans and J. Llusia, 2009 Drought, warming and soil fertilization effects on leaf volatile terpene concentrations in Pinus halepensis and Quercus ilex. Acta Physiologiae Plantarum 31**:** 207.

Bolger, A. M., M. Lohse and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30**:** 2114-2120.

Bonan, G. B., 2008 Forests and climate change: forcings, feedbacks, and the climate benefits of forests. Science 320**:** 1444-1449.

Bondel, K. B., T. Nosenko and W. Stephan, 2018 Signatures of natural selection in abiotic stress-responsive genes of Solanum chilense. R Soc Open Sci 5**:** 171198.

Bowe, L. M., G. Coat and C. W. dePamphilis, 2000 Phylogeny of seed plants based on all three genomic compartments: Extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proceedings of the National Academy of Sciences of the United States of America 97**:** 4092-4097.

Breshears, D. D., N. S. Cobb, P. M. Rich, K. P. Price, C. D. Allen *et al.*, 2005 Regional vegetation die-off in response to global-change-type drought. Proc Natl Acad Sci U S A 102**:** 15144-15148.

Buchanan‐Smith, M., and D. A. Wilhite, 2005 Drought as Hazard: Understanding the Natural and Social Context. Conference Proceedings.

Burgess, D., and M. Freeling, 2014 The most deeply conserved noncoding sequences in plants serve similar functions to those in vertebrates despite large differences in evolutionary rates. Plant Cell 26**:** 946-961.

Busk, P. K., A. B. Jensen and M. Pages, 1997 Regulatory elements in vivo in the promoter of the abscisic acid responsive gene rab17 from maize. Plant Journal 11**:** 1285-1295.

Busk, P. K., and M. Pages, 1997 Protein binding to the abscisic acid-responsive element is independent of VIVIPAROUS1 in vivo. Plant Cell 9**:** 2261-2270.

Busk, P. K., and M. Pages, 1998 Regulation of abscisic acid-induced transcription. Plant Molecular Biology 37**:** 425-435.

Buzeli, R. A. A., J. C. M. Cascardo, L. A. Z. Rodrigues, M. O. Andrade, R. S. Almeida *et al.*, 2002 Tissue-specific regulation of BiP genes: a cis-acting regulatory domain is required for BiP promoter activity in plant meristems. Plant Molecular Biology 50**:** 757-771.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos *et al.*, 2009 BLAST+: architecture and applications. BMC Bioinformatics 10**:** 421.

Casola, C., and T. E. Koralewski, 2018 Pinaceae show elevated rates of gene turnover that are robust to incomplete gene annotation. Plant Journal 95**:** 862-876.

Chang, S. J., J. D. Puryear, M. A. D. L. Dias, E. A. Funkhouser, R. J. Newton *et al.*, 1996 Gene expression under water deficit in loblolly pine (Pinus taeda): Isolation and characterization of cDNA clones. Physiologia Plantarum 97**:** 139-148.

Chaves, M. M., J. Flexas and C. Pinheiro, 2009 Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. Annals of Botany 103**:** 551-560.

Chen, F., X. Zhang, X. Liu and L. Zhang, 2017 Evolutionary Analysis of MIKC(c)-Type MADS-Box Genes in Gymnosperms and Angiosperms. Front Plant Sci 8**:** 895.

Cohen, D., M.-B. Bogeat-Triboulot, E. Tisserant, S. Balzergue, M.-L. Martin-Magniette *et al.*, 2010 Comparative transcriptomics of drought responses in Populus: a meta-analysis of genome-wide expression profiling in mature leaves and root apices across two genotypes. BMC Genom 11**:** 630.

Colquhoun, D., 2014 An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci 1**:** 140216.

Cserhati, M., 2015 Motif content comparison between monocot and dicot species. Genomics Data 3**:** 128-136.

Cui, L., K. Feng, M. Wang, M. Wang, P. Deng *et al.*, 2016 Genome-wide identification, phylogeny and expression analysis of AP2/ERF transcription factors family in Brachypodium distachyon. BMC genomics 17**:** 636.

Cumbie, W. P., A. Eckert, J. Wegrzyn, R. Whetten, D. Neale *et al.*, 2011 Association genetics of carbon isotope discrimination, height and foliar nitrogen in a natural population of Pinus taeda L. Heredity (Edinb) 107**:** 105-114.

Danino, Y. M., D. Even, D. Ideses and T. Juven-Gershon, 2015 The core promoter: At the heart of gene expression. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1849**:** 1116-1131.

De La Torre, A. R., D. Puiu, M. W. Crepeau, K. Stevens, S. L. Salzberg *et al.*, 2018 Genomic architecture of complex traits in loblolly pine. New Phytol.

De La Torre, A. R., B. Wilhite and D. B. Neale, 2019 Environmental Genome-Wide Association Reveals Climate Adaptation Is Shaped by Subtle to Moderate Allele Frequency Shifts in Loblolly Pine. Genome Biol Evol 11**:** 2976-2989.

Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra *et al.*, 2010 Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). Genetics 185**:** 969-982.

Eckert, A. J., J. L. Wegrzyn, J. D. Liechty, J. M. Lee, W. P. Cumbie *et al.*, 2013 The Evolutionary Genetics of the Genes Underlying Phenotypic Associations for Loblolly Pine (Pinus taeda, Pinaceae). Genetics 195**:** 1353-+.

Ersoz, E. S., M. H. Wright, S. C. Gonzalez-Martinez, C. H. Langley and D. B. Neale, 2010 Evolution of disease response genes in loblolly pine: insights from candidate genes. PLoS One 5**:** e14234.

Eveno, E., C. Collada, M. A. Guevara, V. Leger, A. Soto *et al.*, 2008 "Contrasting patterns of selection at Pinus pinaster Ait. Drought stress candidate genes as revealed by genetic differentiation analyses". Mol Biol Evol 25**:** 417-437.

Ezcurra, I., M. Ellerstrom, P. Wycliffe, K. Stalberg and L. Rask, 1999 Interaction between composite elements in the napA promoter: both the B-box ABA-responsive complex and the RY/G complex are necessary for seed-specific expression. Plant Molecular Biology 40**:** 699-709.

Fauteux, F., M. Blanchette and M. V. Stromvik, 2008 Seeder: discriminative seeding DNA motif discovery. Bioinformatics 24**:** 2303-2307.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155**:** 1405-1413.

Fickett, J. W., and A. G. Hatzigeorgiou, 1997 Eukaryotic promoter recognition. Genome Res 7**:** 861-878.

Flanagan, S. P., and A. G. Jones, 2017 Constraints on the FST-Heterozygosity Outlier Approach. J Hered 108**:** 561-573.

Fox, H., A. Doron-Faigenboim, G. Kelly, R. Bourstein, Z. Attia *et al.*, 2018a Transcriptome analysis of Pinus halepensis under drought stress and during recovery. Tree Physiology 38**:** 423-441.

Fox, H., A. Doron-Faigenboim, G. Kelly, R. Bourstein, Z. Attia *et al.*, 2018b Transcriptome analysis of Pinus halepensis under drought stress and during recovery. Tree Physiol 38**:** 423-441.

Fu, Y. X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics 133**:** 693-709.

Fujimoto, S. Y., M. Ohta, A. Usui, H. Shinshi and M. Ohme-Takagi, 2000 Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. Plant Cell 12**:** 393-404.

Golldack, D., C. Li, H. Mohan and N. Probst, 2014 Tolerance to drought and salt stress in plants: unraveling the signaling networks. Frontiers in Plant Science 5.

Gonzalez-Ibeas, D., P. J. Martinez-Garcia, R. A. Famula, A. Delfino-Mix, K. A. Stevens *et al.*, 2016 Assessing the Gene Content of the Megagenome: Sugar Pine (Pinus lambertiana). G3-Genes Genomes Genetics 6**:** 3787-3802.

Gonzalez-Martinez, S. C., E. Ersoz, G. R. Brown, N. C. Wheeler and D. B. Neale, 2006 DNA sequence variation and selection of tag single-nucleotide polymorphisms at candidate genes for drought-stress response in Pinus taeda L. Genetics 172**:** 1915-1926.

Gonzalez-Martinez, S. C., D. Huber, E. Ersoz, J. M. Davis and D. B. Neale, 2008 Association genetics in Pinus taeda L. II. Carbon isotope discrimination. Heredity (Edinb) 101**:** 19-26.

Gotz, S., J. M. Garcia-Gomez, J. Terol, T. D. Williams, S. H. Nagaraj *et al.*, 2008 High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36**:** 3420-3435.

Gould, B. A., J. D. Palacio-Mejia, J. Jenkins, S. Mamidi, K. Barry *et al.*, 2018 Population genomics and climate adaptation of a C4 perennial grass, Panicum hallii (Poaceae). BMC Genomics 19**:** 792.

Graham, J. H., J. J. Duda, M. L. Brown, S. Kitchen, J. M. Emlen *et al.*, 2012 The effects of drought and disturbance on the growth and developmental instability of loblolly pine (Pinus taeda L.). Ecological Indicators 20**:** 143-150.

Greis, D. N. W. a. J. G., 2013 The Southern Forest Futures Project: Technical Report. United States Department of Agriculture, Forest Service.

Grissom, J. E., and R.C. Schmidtling, 1997 Genetic diversity of loblolly pine grown in managed plantations: evidence of differential response to climatic events [Abstract]. In Proceedings of the 24th southern forest tree improvement conference.

Grivet, D., K. Avia, A. Vaattovaara, A. J. Eckert, D. B. Neale *et al.*, 2017 High rate of adaptive evolution in two widespread European pines. Mol Ecol 26**:** 6857-6870.

Guan, R., Y. Zhao, H. Zhang, G. Fan, X. Liu *et al.*, 2016 Draft genome of the living fossil Ginkgo biloba. Gigascience 5**:** s13742-13016-10154-13741.

Guiltinan, M. J., W. R. Marcotte and R. S. Quatrano, 1990 A Plant Leucine Zipper Protein That Recognizes an Abscisic-Acid Response Element. Science 250**:** 267-271.

Gunther, T., and G. Coop, 2013 Robust identification of local adaptation from allele frequencies. Genetics 195**:** 205-220.

Hadiarto, T., and L. S. P. Tran, 2011 Progress studies of drought-responsive genes in rice. Plant Cell Reports 30**:** 297-310.

Hamberger, B., D. Hall, M. Yuen, C. Oddy, B. Hamberger *et al.*, 2009 Targeted isolation, sequence assembly and characterization of two white spruce (Picea glauca) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defence reveal insights into a conifer genome. BMC Plant Biol 9**:** 106.

Hancock, A. M., G. Alkorta-Aranburu, D. B. Witonsky and A. Di Rienzo, 2010 Adaptations to new environments in humans: the role of subtle allele frequency shifts. Philos Trans R Soc Lond B Biol Sci 365**:** 2459-2468.

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

Hart, A. J., S. Ginzburg, M. S. Xu, C. R. Fisher, N. Rahmatpour *et al.*, 2020 EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. Mol Ecol Resour 20**:** 591-604.

Heath, L. S., N. Ramakrishnan, R. R. Sederoff, R. W. Whetten, B. I. Chevone *et al.*, 2002 Studying the functional genomics of stress responses in loblolly pine with the Expresso microarray experiment management system. Comparative and Functional Genomics 3**:** 226-243.

Hickman, R., M. C. Van Verk, A. J. H. Van Dijken, M. P. Mendes, I. A. Vroegop-Vos *et al.*, 2017 Architecture and Dynamics of the Jasmonic Acid Gene Regulatory Network. Plant Cell 29**:** 2086-2105.

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

Hu, L., Y. Xie, S. Fan, Z. Wang, F. Wang *et al.*, 2018 Comparative analysis of root transcriptome profiles between drought-tolerant and susceptible wheat genotypes in response to water stress. Plant Sci 272**:** 276-293.

Hu, R., B. Wu, H. Zheng, D. Hu, X. Wang *et al.*, 2015 Global Reprogramming of Transcription in Chinese Fir (Cunninghamia lanceolata) during Progressive Drought Stress and after Rewatering. Int J Mol Sci 16**:** 15194-15219.

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics 116**:** 153-159.

Imin, N., M. Nizamidin, T. Wu and B. G. Rolfe, 2007 Factors involved in root formation in Medicago truncatula. J Exp Bot 58**:** 439-451.

Janiak, A., M. Kwasniewski, M. Sowa, K. Gajek, K. Zmuda *et al.*, 2018 No Time to Waste: Transcriptome Study Reveals that Drought Tolerance in Barley May Be Attributed to Stressed-Like Expression Patterns that Exist before the Occurrence of Stress. Frontiers in Plant Science 8.

Janiak, A., M. Kwasniewski and I. Szarejko, 2016 Gene expression regulation in roots under drought. J Exp Bot 67**:** 1003-1014.

Jin, J. P., F. Tian, D. C. Yang, Y. Q. Meng, L. Kong *et al.*, 2017 PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Research 45**:** D1040-D1045.

Johnsen, K. T., Bob; Samuelson, Lisa; Butnor, John; Sampson, David; Sanchez, Felipe; Maier, Chris; McKeand, Steve, 2004 Carbon Sequestration in loblolly pine plantations: Methods, limitations, and research needs for estimating storage pools. In: Gen. Tech. Rep. SRS–75. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. Chapter 32**:** p. 373-381.

Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis *et al.*, 2008 NCBI BLAST: a better web interface. Nucleic Acids Res 36**:** W5-9.

Juven-Gershon, T., J. Y. Hsu, J. W. Theisen and J. T. Kadonaga, 2008 The RNA polymerase II core promoter - the gateway to transcription. Curr Opin Cell Biol 20**:** 253-259.

Karl, T. R., J. M. Melillo and T. C. Peterson, 2009 *Global climate change impacts in the United States*. Cambridge University Press, Cambridge.

Keich, U., and P. A. Pevzner, 2002 Finding motifs in the twilight zone. Bioinformatics 18**:** 1374-1381.

Kim, D., B. Landmead and S. L. Salzberg, 2015 HISAT: a fast spliced aligner with low memory requirements. Nature Methods 12**:** 357-U121.

Kim, D. W., S. H. Lee, S. B. Choi, S. K. Won, Y. K. Heo *et al.*, 2006 Functional conservation of a root hair cell-specific cis-element in angiosperms with different root hair distribution patterns. Plant Cell 18**:** 2958-2970.

Klockow, P. A., C. B. Edgar, G. W. Moore and J. G. Vogel, 2020 Southern Pines Are Resistant to Mortality From an Exceptional Drought in East Texas. Frontiers in Forests and Global Change 3.

Kopylova, E., L. Noe and H. Touzet, 2012 SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28**:** 3211-3217.

Koralewski, T. E., J. E. Brooks and K. V. Krurovsky, 2014 Molecular evolution of drought tolerance and wood strength related candidate genes in loblolly pine (Pinus taeda L.). Silvae Genetica 63**:** 59-66.

Lagrange T, G. S., Yeo HJ, Mache R, 1997 S2F, a leaf-specific trans-acting factor, binds to a novel cis-acting element and differentially activates the RPL21 gene. Plant Cell 9**:** 1469-1479.

Le Corre, V., and A. Kremer, 2012 The genetic differentiation at quantitative trait loci under local adaptation. Molecular Ecology 21**:** 1548-1566.

Lei, Y., C. Yin and C. Li, 2006 Differences in some morphological, physiological, and biochemical responses to drought stress in two contrasting populations of Populus przewalskii. Physiologia Plantarum 127**:** 182-191.

Liao, W., S. Zhao, M. Zhang, K. Dong, Y. Chen *et al.*, 2017 Transcriptome Assembly and Systematic Identification of Novel Cytochrome P450s in Taxus chinensis. Front Plant Sci 8**:** 1468.

Lieberman-Lazarovich, M., C. Yahav, A. Israeli and I. Efroni, 2019 Deep Conservation of cis-Element Variants Regulating Plant Hormonal Responses. Plant Cell 31**:** 2559-2572.

Loopstra, C. A., and R. R. Sederoff, 1995 Xylem-Specific Gene-Expression in Loblolly-Pine. Plant Molecular Biology 27**:** 277-291.

Lorenz, W. W., R. Alba, Y. S. Yu, J. M. Bordeaux, M. Simoes *et al.*, 2011 Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (P. taeda L.). BMC Genomics 12**:** 264.

Lorenz, W. W., F. Sun, C. Liang, D. Kolychev, H. Wang *et al.*, 2006 Water stress-responsive genes in loblolly pine (Pinus taeda) roots identified by analyses of expressed sequence tag libraries. Tree Physiol 26**:** 1-16.

Lou, Q., L. Chen, H. Mei, K. Xu, H. Wei *et al.*, 2017 Root Transcriptomic Analysis Revealing the Importance of Energy Metabolism to the Development of Deep Roots in Rice (Oryza sativa L.). Front Plant Sci 8**:** 1314.

Love, M. I., W. Huber and S. Anders, 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15.

Lu, M., K. V. Krutovsky, C. D. Nelson, T. E. Koralewski, T. D. Byram *et al.*, 2016 Exome genotyping, linkage disequilibrium and population structure in loblolly pine (Pinus taeda L.). BMC Genomics 17**:** 730.

Lu, M., C. A. Loopstra and K. V. Krutovsky, 2019 Detecting the genetic basis of local adaptation in loblolly pine (Pinus taeda L.) using whole exome-wide genotyping and an integrative landscape genomics analysis approach. Ecol Evol 9**:** 6798-6809.

Lu, M. M., K. V. Krutovsky, C. D. Nelson, J. B. West, N. A. Reilly *et al.*, 2017 Association genetics of growth and adaptive traits in loblolly pine (Pinus taeda L.) using whole-exome-discovered polymorphisms. Tree Genetics & Genomes 13.

Lu, M. M., C. M. Seeve, C. A. Loopstra and K. V. Krutovsky, 2018 Exploring the genetic basis of gene transcript abundance and metabolite levels in loblolly pine (Pinus taeda L.) using association mapping and network construction. Bmc Genetics 19.

Maeso, I., M. Irimia, J. J. Tena, F. Casares and J. L. Gomez-Skarmeta, 2013 Deep conservation of cis-regulatory elements in metazoans. Philos Trans R Soc Lond B Biol Sci 368**:** 20130020.

Magwanga, R. O., J. N. Kirungu, P. Lu, X. Yang, Q. Dong *et al.*, 2019 Genome wide identification of the trihelix transcription factors and overexpression of Gh_A05G2067 (GT-2), a novel gene contributing to increased drought and salt stresses tolerance in cotton. Physiol Plant 167**:** 447-464.

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

McKay, J. K., J. H. Richards and T. Mitchell-Olds, 2003 Genetics of drought adaptation in Arabidopsis thaliana: I. Pleiotropy contributes to genetic correlations among ecological traits. Molecular Ecology 12**:** 1137-1151.

McKiernan, A. B., M. J. Hovenden, T. J. Brodribb, B. M. Potts, N. W. Davies *et al.*, 2014 Effect of limited water availability on foliar plant secondary metabolites of two Eucalyptus species. Environmental and Experimental Botany 105**:** 55-64.

Mia, M. S., H. Liu, X. Wang, C. Zhang and G. Yan, 2020 Root transcriptome profiling of contrasting wheat genotypes provides an insight to their adaptive strategies to water deficit. Sci Rep 10**:** 4854.

Michael, R., A. Ranjan, R. S. Kumar, P. K. Pathak and P. K. Trivedi, 2020 Light-regulated expression of terpene synthase gene, AtTPS03, is controlled by the bZIP transcription factor, HY5, in Arabidopsis thaliana. Biochemical and Biophysical Research Communications 529**:** 437-443.

Michael, T. P., T. C. Mockler, G. Breton, C. McEntee, A. Byer *et al.*, 2008 Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genet 4**:** e14.

Mizoi, J., K. Shinozaki and K. Yamaguchi-Shinozaki, 2012 AP2/ERF family transcription factors in plant abiotic stress responses. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1819**:** 86-96.

Moran, E., J. Lauder, C. Musser, A. Stathos and M. Shu, 2017 The genetics of drought tolerance in conifers. New Phytologist 216**:** 1034-1048.

Mosca, E., F. Cruz, J. Gómez-Garrido, L. Bianco, C. Rellstab *et al.*, 2019 A reference genome sequence for the European silver fir (Abies alba Mill.): a community-generated genomic resource. G3: Genes, Genomes, Genetics 9**:** 2039-2049.

Mu, M., X. K. Lu, J. J. Wang, D. L. Wang, Z. J. Yin *et al.*, 2016 Genome-wide Identification and analysis of the stress-resistance function of the TPS (Trehalose-6-Phosphate Synthase) gene family in cotton. BMC Genet 17**:** 54.

Muller, T., I. Ensminger and K. J. Schmid, 2012 A catalogue of putative unique transcripts from Douglas-fir (Pseudotsuga menziesii) based on 454 transcriptome sequencing of genetically diverse, drought stressed seedlings. BMC Genomics 13**:** 673.

Nakabayashi, R., K. Yonekura‐Sakakibara, K. Urano, M. Suzuki, Y. Yamada *et al.*, 2014 Enhancement of oxidative and drought tolerance in Arabidopsis by overaccumulation of antioxidant flavonoids. The Plant Journal 77**:** 367-379.

Neale, D. B., and A. Kremer, 2011 Forest tree genomics: growing resources and applications. Nat Rev Genet 12**:** 111-122.

Neale, D. B., P. E. McGuire, N. C. Wheeler, K. A. Stevens, M. W. Crepeau *et al.*, 2017 The Douglas-Fir Genome Sequence Reveals Specialization of the Photosynthetic Apparatus in Pinaceae. G3 (Bethesda) 7**:** 3157-3167.

Noctor, G., A. Mhamdi and C. H. Foyer, 2014 The roles of reactive oxygen metabolism in drought: not so cut and dried. Plant Physiol 164**:** 1636-1648.

Nuruzzaman, M., A. M. Sharoni and S. Kikuchi, 2013 Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. Front Microbiol 4**:** 248.

Nystedt, B., N. R. Street, A. Wetterbom, A. Zuccolo, Y.-C. Lin *et al.*, 2013 The Norway spruce genome sequence and conifer genome evolution. Nature 497**:** 579-584.

Oikkonen, L., and S. Lise, 2017 Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. Wellcome Open Res 2**:** 6.

Osakabe, Y., K. Osakabe, K. Shinozaki and L. S. Tran, 2014 Response of plants to water stress. Front Plant Sci 5**:** 86.

Palle, S. R., C. M. Seeve, A. J. Eckert, W. P. Cumbie, B. Goldfarb *et al.*, 2011 Natural variation in expression of genes involved in xylem development in loblolly pine (Pinus taeda L.). Tree Genetics & Genomes 7**:** 193-206.

Pascual, M. B., M. T. Llebres, B. Craven-Bartle, R. A. Canas, F. M. Canovas *et al.*, 2018 PpNAC1, a main regulator of phenylalanine biosynthesis and utilization in maritime pine. Plant Biotechnol J 16**:** 1094-1104.

Peleg, Z., Y. Saranga, T. Krugman, S. Abbo, E. Nevo *et al.*, 2008 Allelic diversity associated with aridity gradient in wild emmer wheat populations. Plant Cell and Environment 31**:** 39-49.

Perdiguero, P., C. Barbero Mdel, M. T. Cervera, C. Collada and A. Soto, 2013 Molecular response to water stress in two contrasting Mediterranean pines (Pinus pinaster and Pinus pinea). Plant Physiol Biochem 67**:** 199-208.

Pertea, M., D. Kim, G. M. Pertea, J. T. Leek and S. L. Salzberg, 2016 Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols 11**:** 1650-1667.

Pertea, M., G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell *et al.*, 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 33**:** 290-+.

Peviani, A., J. Lastdrager, J. Hanson and B. Snel, 2016 The phylogeny of C/S1 bZIP transcription factors reveals a shared algal ancestry and the pre-angiosperm translational regulation of S1 transcripts. Sci Rep 6**:** 30444.

Pfeifer, B., U. Wittelsburger, S. E. Ramos-Onsins and M. J. Lercher, 2014 PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol 31**:** 1929-1936.

Pratima Devkota, S. A. E., Lori G. Eckhardt, 2018 The Impact of Drought and Vascular-Inhabiting Pathogen Invasion in Pinus taeda Health. International Journal of Forestry Research 2018.

Prisley, C. L. V. a. S. P., 2019 FACTORS AFFECTING SITE PRODUCTIVITY OF LOBLOLLY PINE PLANTATIONS ACROSS THE SOUTHEASTERN UNITED STATES. Conference: Proceedings of the 5th Southern Forestry and Natural Resources GIS Conference.

Pritchard, J. K., J. K. Pickrell and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20**:** R208-215.

Prunier, J., J. P. Verta and J. J. MacKay, 2016 Conifer genomics and adaptation: at the crossroads of genetic diversity and genome function. New Phytologist 209**:** 44-62.

Qiu, Q., T. Ma, Q. Hu, B. Liu, Y. Wu *et al.*, 2011 Genome-scale transcriptome analysis of the desert poplar, Populus euphratica. Tree physiology 31**:** 452-461.

Quesada, T., V. Gopal, W. P. Cumbie, A. J. Eckert, J. L. Wegrzyn *et al.*, 2010 Association mapping of quantitative disease resistance in a natural population of loblolly pine (Pinus taeda L.). Genetics 186**:** 677-686.

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26**:** 841-842.

reviewed in Signor, S. A., and S. V. Nuzhdin, 2018 The Evolution of Gene Expression in cis and trans. Trends in Genetics 34**:** 532-544.

Rigal, A., Y. S. Yordanov, I. Perrone, A. Karlberg, E. Tisserant *et al.*, 2012 The AINTEGUMENTA LIKE1 homeotic transcription factor PtAIL1 controls the formation of adventitious root primordia in poplar. Plant Physiol 160**:** 1996-2006.

Roy, A. L., and D. S. Singer, 2015 Core promoters in transcription: old problem, new insights. Trends in biochemical sciences 40**:** 165-171.

Ruiz Daniels, R., R. S. Taylor, M. J. Serra-Varela, G. G. Vendramin, S. C. Gonzalez-Martinez *et al.*, 2018 Inferring selection in instances of long-range colonization: The Aleppo pine (Pinus halepensis) in the Mediterranean Basin. Mol Ecol.

Sanchez-Munoz, R., M. Bonfill, R. M. Cusido, J. Palazon and E. Moyano, 2018 Advances in the Regulation of In Vitro Paclitaxel Production: Methylation of a Y-Patch Promoter Region Alters BAPT Gene Expression in Taxus Cell Cultures. Plant Cell Physiol 59**:** 2255-2267.

Schmidtling, R. C., 2001 Southern Pine Seed Sources, pp.

Seki, M., M. Narusaka, H. Abe, M. Kasuga, K. Yamaguchi-Shinozaki *et al.*, 2001 Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. Plant Cell 13**:** 61-72.

Silva, C. S., S. Puranik, A. Round, M. Brennich, A. Jourdain *et al.*, 2015 Evolution of the Plant Reproduction Master Regulators LFY and the MADS Transcription Factors: The Role of Protein Structure in the Evolutionary Development of the Flower. Front Plant Sci 6**:** 1193.

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

Singh, M., J. Kumar, S. Singh, V. P. Singh and S. M. Prasad, 2015 Roles of osmoprotectants in improving salinity and drought tolerance in plants: a review. Reviews in Environmental Science and Bio/Technology 14**:** 407-426.

Spensley, M., J. Y. Kim, E. Picot, J. Reid, S. Ott *et al.*, 2009 Evolutionarily conserved regulatory motifs in the promoter of the Arabidopsis clock gene LATE ELONGATED HYPOCOTYL. Plant Cell 21**:** 2606-2623.

Sperry, J. S., U. G. Hacke, R. Oren and J. P. Comstock, 2002 Water deficits and hydraulic limits to leaf water supply. Plant Cell and Environment 25**:** 251-263.

Steane, D. A., B. M. Potts, E. McLean, L. Collins, S. M. Prober *et al.*, 2015 Genome-wide scans reveal cryptic population structure in a dry-adapted eucalypt. Tree Genetics & Genomes 11.

Steane, D. A., B. M. Potts, E. McLean, S. M. Prober, W. D. Stock *et al.*, 2014 Genome-wide scans detect adaptation to aridity in a widespread forest tree species. Molecular Ecology 23**:** 2500-2513.

Supek, F., M. Bošnjak, N. Škunca and T. Šmuc, 2011 REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one 6**:** e21800.

Szklarczyk, D., A. L. Gable, D. Lyon, A. Junge, S. Wyder *et al.*, 2019 STRING v11: protein-protein association networks with increased coverage, supporting

functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47**:** D607-D613.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123**:** 585-595.

Tang, S. Y. Y., A. Lomsadze and M. Borodovsky, 2015 Identification of protein coding regions in RNA transcripts. Nucleic Acids Research 43.

Tang, T., A. M. Yu, P. Li, H. Yang, G. J. Liu *et al.*, 2016 Sequence analysis of the Hsp70 family in moss and evaluation of their functions in abiotic stress responses. Scientific Reports 6.

Terzaghi, W. B., and A. R. Cashmore, 1995 Light-Regulated Transcription. Annual Review of Plant Physiology and Plant Molecular Biology 46**:** 445-474.

Timmerhaus, G., S. T. Hanke, K. Buchta and S. A. Rensing, 2011 Prediction and validation of promoters involved in the abscisic acid response in Physcomitrella patens. Mol Plant 4**:** 713-729.

Tran, L. S. P., K. Nakashima, Y. Sakuma, S. D. Simpson, Y. Fujita *et al.*, 2004 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 16**:** 2481-2498.

Tripathi, P., R. C. Rabara and P. J. Rushton, 2014 A systems biology perspective on the role of WRKY transcription factors in drought responses in plants. Planta 239**:** 255-266.

Turco, G., J. C. Schnable, B. Pedersen and M. Freeling, 2013 Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. Front Plant Sci 4**:** 170.

van Mantgem, P. J., and N. L. Stephenson, 2007 Apparent climatically induced increase of tree mortality rates in a temperate forest. Ecol Lett 10**:** 909-916.

Vandepoele, K., M. Quimbaya, T. Casneuf, L. De Veylder and Y. Van de Peer, 2009 Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol 150**:** 535-546.

Vikas Shalibhadra Trishla, S. M., Prasanna Boyidi, Padmaja Gudipalli, Pulugurtha Bharadwaja Kirti, 2019 Characterization of a vascular bundle localizing Gossypium hirsutum NAC4 transcription factor promoter for its role in environmental stress responses.

Villar, D., P. Flicek and D. T. Odom, 2014 Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat Rev Genet 15**:** 221-233.

Vogt, T., 2010 Phenylpropanoid biosynthesis. Molecular plant 3**:** 2-20.

Vuosku, J., K. Karppinen, R. Muilu-Makela, T. Kusano, G. H. M. Sagor *et al.*, 2018 Scots pine aminopropyltransferases shed new light on evolution of the polyamine biosynthesis pathway in seed plants. Ann Bot 121**:** 1243-1256.

W. Brad Smith, P. D. M., Charles H. Perry, Scott A. Pugh, 2007 Forest Resources of the United States.

Wahlenberg, W. G., 1960 Loblolly pine, its use, ecology, regeneration, protection, growth and management. Durham, NC: Duke University, School of Forestry 603 p.

Wan, T., Z.-M. Liu, L.-F. Li, A. R. Leitch, I. J. Leitch *et al.*, 2018 A genome for gnetophytes and early evolution of seed plants. Nature Plants 4**:** 82-89.

Wang, H., K. Li, X. Sun, Y. Xie, X. Han *et al.*, 2019 Isolation and characterization of larch BABY BOOM2 and its regulation of adventitious root development. Gene 690**:** 90-98.

Warren, R. L., C. I. Keeling, M. M. S. Yuen, A. Raymond, G. A. Taylor *et al.*, 2015 Improved white spruce (Picea glauca) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. The Plant Journal 83**:** 189-212.

Watkinson, J. I., A. A. Sioson, C. Vasquez-Robinet, M. Shukla, D. Kumar *et al.*, 2003 Photosynthetic acclimation is reflected in specific patterns of gene expression in drought-stressed loblolly pine. Plant Physiology 133**:** 1702-1716.

Wegrzyn, J. L., J. D. Liechty, K. A. Stevens, L. S. Wu, C. A. Loopstra *et al.*, 2014 Unique Features of the Loblolly Pine (Pinus taeda L.) Megagenome Revealed Through Sequence Annotation. Genetics 196**:** 891-+.

Whitlock, M. C., and K. E. Lotterhos, 2015 Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F(ST). Am Nat 186 Suppl 1**:** S24-36.

Wittkopp, P. J., and G. Kalay, 2012 Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature Reviews Genetics 13**:** 59-69.

Wu, Z., X. Xu, W. Xiong, P. Wu, Y. Chen *et al.*, 2015 Genome-Wide Analysis of the NAC Gene Family in Physic Nut (Jatropha curcas L.). PLoS One 10**:** e0131890.

Xie, Z., T. M. Nolan, H. Jiang and Y. Yin, 2019 AP2/ERF Transcription Factor Regulatory Networks in Hormone and Abiotic Stress Responses in Arabidopsis. Front Plant Sci 10**:** 228.

Xie, Z. M., H. F. Zou, G. Lei, W. Wei, Q. Y. Zhou *et al.*, 2009 Soybean Trihelix transcription factors GmGT-2A and GmGT-2B improve plant tolerance to abiotic stresses in transgenic Arabidopsis. PLoS One 4**:** e6898.

Yamamoto, Y. Y., H. Ichida, M. Matsui, J. Obokata, T. Sakurai *et al.*, 2007 Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics 8**:** 67.

Yanfang, Y., Z. Kaikai, Y. Liying, L. Xing, W. Ying *et al.*, 2018 Identification and characterization of MYC transcription factors in Taxus sp. Gene 675**:** 1-8.

Ye, J. W., W. N. Bai, L. Bao, T. M. Wang, H. F. Wang *et al.*, 2017 Sharp genetic discontinuity in the aridity-sensitive Lindera obtusiloba (Lauraceae): solid evidence supporting the Tertiary floral subdivision in East Asia. Journal of Biogeography 44**:** 2082-2095.

Yevtushenko, D. P., and S. Misra, 2018 Spatiotemporal activities of Douglas-fir BiP Pro1 promoter in transgenic potato. Planta 248**:** 1569-1579.

Yin, T. Z., G. Pan, H. Liu, J. Wu, Y. P. Li *et al.*, 2012 The chloroplast ribosomal protein L21 gene is essential for plastid development and embryogenesis in Arabidopsis. Planta 235**:** 907-921.

You, J., Y. Zhang, A. Liu, D. Li, X. Wang *et al.*, 2019 Transcriptomic and metabolomic profiling of drought-tolerant and susceptible sesame genotypes in response to drought stress. BMC plant biology 19**:** 1-16.

Yu, C., L. Song, J. Song, B. Ouyang, L. Guo *et al.*, 2018 ShCIGT, a Trihelix family gene, mediates cold and drought tolerance by interacting with SnRK1 in tomato. Plant Sci 270**:** 140-149.

Zhang, X., J. Pang, X. Ma, Z. Zhang, Y. He *et al.*, 2019 Multivariate analyses of root phenotype and dynamic transcriptome underscore valuable root traits and water-deficit responsive gene networks in maize. Plant Direct 3**:** 1-18.

Zolotarov, Y., and M. Stromvik, 2015 De Novo Regulatory Motif Discovery Identifies Significant Motifs in Promoters of Five Classes of Plant Dehydrin Genes. Plos One 10.