UNIVERSITY OF CALIFORNIA, MERCED

Association genetics of drought tolerance in ponderosa pine (*Pinus ponderosa*)

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

In

Environmental Systems

by

Mengjun Shu

Committee in Charge:

Professor Stephen C. Hart, Chair
Professor Emily V. Moran, Research Advisor
Professor David B. Neale
Professor Jason Sexton

2020

The Dissertation of Mengjun Shu, titled Association genetics of drought tolerance in ponderosa pine (*Pinus ponderosa*), is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Professor David B. Neale	Date
Professor Jason Sexton	Date
Professor Emily V. Moran, Advisor	Date
Professor Stephen C. Hart, Chair	Date

University of California, Merced
2020

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to acknowledge all funding programs that make it possible for me to conduct research and earn my Ph.D. degree. Specifically, I want to thank the Forest Service's Pacific Southwest Regional Genetic Resources Program for allowing us to sample needles and seeds from their seed orchard. The sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. For SNP identification, the calculation was done in the MERCED computer cluster at UC Merced (supported by NSF Award ACI-1429783) and Bridges in Pittsburgh Supercomputing Center under Extreme Science and Engineering Discovery Environment (XSEDE Project # TG-DEB190006), supported by NSF Award ACI-1548562. My research was also partially supported by funding from a Graduate Student Opportunity Program Fellowship at UCM and Environmental Systems (ES) travel and summer support.

Next, I want to express my deepest gratitude to my advisor Emily Moran. She was passionate and "brave" enough to recruit and fund me as a PhD student in her group despite the fact that I came from a less known foreign background and that did not have a Master degree. That was how our adventure started. With her considerate guidance and consistent financial and emotional support, I became able to address several challenging problems and enjoy doing research. Also, as a successful female scientist, she's always been a role model for me. Being her Ph.D. student, TA, collaborator, I never stop learning something new from her. I also appreciate the time-spending and insightful comments from all my committee members. The regular meetings and reports kept me well on track of my research and earning my degree. Additionally, Steve Hart is very generous in allowing me to borrow the expensive apparatus in his lab whenever I need; Jay Sexton is always available to ease my tension with his humor; David Neale provides huge networking resources to help me in clarifying the details in data analysis.

I would like to express my gratitude and appreciation to Ecologists Anonymous, a group of like-minded ecology students from the ES and QSB graduate groups. The bi-weekly meetings in Ecologists Anonymous was extremely helpful in improving oral English, preparing for qualifying exam, writing fellowship and grant applications in my early stage and making me overcome the cultural gap and feel like a part of an ecology graduate student community. I personally thank Jeffery Lauder, Robert Boria, Nate Fox, Danaan Deneve, Jackie Shay, Lillie Pennington, Laura Van Vranken, Daniel Toews, and Kinsey Brock for their precious feedback and friendship.

Lastly, I'd like to thank my families for their caring and love. I would not be able to study abroad without their support. And for my partner, Peizhi Mai, for his love, patience and being supportive in any situation.

Mengjun Shu
PhD candidate
School of Engineering, University of California, Merced
5200 North Lake Rd, Merced, CA, 95343
mshu@ucmerced.edu

EDUCATION

Ph.D. Environmental Systems Summer 2020 (Anticipated)
University of California, Merced
Thesis title: "Association genetics of drought tolerance in ponderosa pine (*Pinus ponderosa*)"

B.S. Geography science Spring 2013
Sun Yat-sen University (SYSU), Guangzhou, China
Thesis title: "Diversity and dispersion of climbing plants in Southeast Asia"

RESEARCH EXPERIENCES

Graduate Student Researcher Aug 2015 – Present
University of California, Merced CA

Lab assistant Sep 2014 – May 2015
Prof. Liang Hu's lab in SYSU, Guangzhou, China

Lab volunteer Jan 2014 – July 2014
Prof. Ingrid Parker's lab in University of California, Santa Cruz, CA

Lab assistant July 2013 – Dec 2013
Research Center of Forest Ecosystem in Tropics and Subtropics, Guangzhou, China

Undergraduate student researcher Sep 2011 – June 2013
Prof. Liang Hu's lab in SYSU, Guangzhou, China

AWARDS & FELLOWSHIPS

Graduate Student Opportunity Program Fellowship Fall 2018 – Summer 2019
Environmental Systems, UC Merced; $42,747.29

Summer Travel Fellowship Summer 2018
Environmental Systems, UC Merced; $1000

Peer Mentor Fellowship Fall 2017– Spring 2018

Environmental Systems, UC Merced; $500

Graduate Student Summer Fellowship Summer 2017
Environmental Systems, UC Merced; $7500
UC Conservation Genomics Consortium Catalyst Grant Spring 2017
University of California, Los Angeles; $250

Undergraduate Honors Thesis June 2013
Based on written thesis and thesis defense; SYSU

Third-level Scholarship Fall 2011
SYSU; CNY2000

SUPERCOMPUTING RESOURCES

The XSEDE research project July 2019 – June 2020
Title: Association genetics of drought tolerance in ponderosa pine (*Pinus ponderosa*)
Resources: 30,000 SUs in Bridges Large and 4000 GB in Bridges Storage ($20,340)

The XSEDE startup project July 2018 – June 2019
Title: Association genetics of drought tolerance in ponderosa pine (*Pinus ponderosa*)
Resources: 9000 SUs in Bridges Large and 2000 GB in Bridges Storage

WORKSHOP

UC Merced Dissertation Writing Boot Camp 2017-2019
UC Merced Yosemite Software Carpentry Workshop August, 2017
22nd Summer Institute in Statistical Genetics, UW July, 2017
UCLA/La Kretz Workshop in Conservation Genomics March, 2017

TALKS/POSTERS

M. Shu, 2019, "Testing pipelines for genome-wide SNP calling from Genotyping-by-Sequencing data for *Pinus ponderosa*", IUFRO Tree Biotechnology Meeting, Raleigh, NC – poster

M. Shu, E. V. Moran, **2018**, "Responses to water and soil conditions in ponderosa pine seedlings", Ecological Society of America, New Orleans, LA – oral presentation

TEACHING EXPERIENCE

Nutrition. *Teaching Assistant*. UC Merced. Spring 2020.
Evolution. *Teaching Assistant*. UC Merced. Summer & Fall 2019.
Global Change Biology. *Guest lecture*. UC Merced. Fall 2018

Biodiversity and Conservation. *Teaching Assistant*. UC Merced. Spring & Fall 2017
Plant Biology. *Teaching Assistant*. UC Merced. Spring 2018
Introductory Biology labs. *Teaching Assistant*. UC Merced. Spring 2016 & Summer 2017

PUBLICATIONS

M. Shu & E.V. Moran. **In prep**, Genome-wide SNP calling of *Pinus ponderosa* from genotyping-by-sequencing (GBS) data using different pipelines. *Plant Methods*.

Moran, E.V., J. Lauder, C. Musser, A. Stathos, **M. Shu**. **2018.** Genetics of drought tolerance in conifers and its implications for adaptation to climate change. *New Phytologist*. Tansley Reviews. 216:1034–1048

Abstract

Association genetics of drought tolerance in ponderosa pine (*Pinus ponderosa*)

by

Mengjun Shu

Doctor of Philosophy in Environmental Systems

University of California, Merced
2020

Professor Stephen C. Hart, Chair
Professor Emily V. Moran, Research Advisor

Drought stress is a major cause of tree mortality in Mediterranean coniferous forests. This study aims to investigate the genetics of drought tolerance in ponderosa pine (*Pinus ponderosa*), a highly valuable species in the western United States. Genotype-to-environment (G2E) association investigates the statistical association between genetic variation at individual loci and the environment, while genotype-to-phenotype (G2P) association identifies loci linked to a particular phenotype by correlating genotypes at SNPs with the variation in certain traits. By combining G2E and G2P association genetics, this study can identify both the loci and traits that may explain variation in drought tolerance in this pine species. Single Nucleotide Polymorphism (SNP) markers have rapidly gained popularity due to their abundance in most genomes and their amenability to high-throughput genotyping techniques. Genotyping-by-sequencing (GBS) has been demonstrated to be a robust and cost-effective genotyping method. We first compared the performance of four GBS bioinformatics pipelines, two of which require a reference genome (TASSEL-GBS V2 and Stacks), two of which are *de novo* pipelines (UNEAK and Stacks), on this large-genome non-model organism. Stacks with a reference genome produced the highest number of SNPs with lowest proportion of paralogs. Over 4 million SNPs were identified with 223 ponderosa pine individuals using this method and the reference genome of loblolly pine (*Pinus taeda*). Then I ran a G2E analysis with these SNPs and five chosen climatic variables using LFMM2, which controls for the effects of demographic processes and population structure on the distribution of genetic variation. I found 213 SNPs strongly associated with mean maximum temperature of summer, 335 with mean minimum temperature of winter, 1798 with April 1st snow pack, and 120 SNPs with mean climatic water deficit. Protein functions linked to associated SNPs include ubiquitination, the abscisic acid (ABA) signaling pathway, cell division or growth of

roots or shoots, cell wall organization, seed dormancy. The G2P analysis was carried out based on greenhouse experiment data. Seeds from 48 genotyped mother trees were planted in both dry and wet treatments. Eight phenotypic traits were measured during or after the greenhouse experiment. Six were drought-responsive, including root length, root-shoot dry mass ratio, stomata density on adaxial side, and number of stomatal rows on abaxial side (all higher in dry treatment), as well as shoot weight and height growth (lower). I found 153 SNPs strongly associated with root length, 80 with shoot weight, 145 with height growth, 42 with adaxial stomatal density, 85 with abaxial stomatal rows, and 1530 with root-to-shoot ratio. The identified SNPs reside in genes with a wide variety of functions, including ubiquitination, abscisic acid (ABA) signaling pathway, cell division or growth of roots or shoots, cell wall organization, which overlap with most of the identified protein functions in the G2E analysis. Potentially, future studies can develop molecular tools based on the associated genetic markers to assist breeders and gene resource managers in developing and managing adapted populations.

Chapter 1:
Introduction

1.1 Climate change and forest health

High rates of tree mortality caused by drought are occurring in many regions of the world (Loarie et al. 2009, Pereira et al. 2010, Allen et al. 2010). Global climate change over the next century is predicted to result in higher evaporative demand, changing amount and type of precipitation, and earlier snowmelt, which will result in warmer, longer, and more frequent drought in already arid and semi-arid environments (Ryan 2011, IPCC 2014, Pachauri et al. 2014). Risks from hot drought are particularly relevant to California, the Mediterranean climate of which is characterized by winter precipitation and a hot dry summer (Royce and Barbour 2001, Giorgi and Lionello 2008). For example, the California drought of 2012-2016 was the most severe drought in the past millennium, leading to an estimated 130 million standing dead trees in the Sierra Nevada and negatively impacting the sustainability of conifer forests (Williams et al. 2015, Schultz 2017, Fettig et al. 2019).

Given the projected increases in temperature due to climate change, California's 2012–2016 drought may represent an increasingly common condition in which warmer temperatures coincide with periodically occurring dry years (Berg and Hall 2015). Such conditions would lead to increased water stress in Mediterranean trees species under changing climate (Allen et al. 2010, Williams et al. 2015). Drought affects tree responses from the molecular level to the forest stand level, and is a major cause of tree mortality (Newton et al. 1991, van Mantgem et al. 2009, Allen et al. 2010, Hamanishi and Campbell 2011, Goulden and Bales 2014). Due to their long life span and lack of mobility, trees are especially susceptible to the effects of climate. Forest trees play a critical role in terrestrial ecosystems, offering major ecological benefits in terms of carbon fixation, soil retention, and wildlife habitats. Understanding how forest trees respond to the drought is critical for forest management and sustainability.

1.2 Drought tolerance and the genes underlying it

The ability of a tree to survive or grow under dry conditions is defined as drought tolerance. There are three categories of drought tolerance. Drought avoidance strategies reduce exposure to drought stress by adjusting physiological traits, such as by growing deep roots or controlling stomatal openings (McDowell et al. 2008). Drought resistance is the ability of a tree to resist growth loss due to drought (Montwé et al. 2016), while drought resilience quantifies how quickly a tree can recover to normal growth when conditions improve (Lloret et al. 2011, Eilmann and Rigling 2012).

A long history of studies in forestry have clearly demonstrated the existence of local adaptation in tree populations (Langlet 1971, Ying and Liang 1994, Kitzmiller 2005, Wright 2007). However, locally adapted tree populations with long life cycles may become maladapted if climate-induced environmental shifts outpace range shifts, plastic responses, or evolutionary adaptation (Aitken et al. 2008, Anderson et al. 2012, Alberto et al. 2013). Adaptive genetic variation, which represents adaptation within species and

populations via changes in allele frequencies or genotypic recombination, is therefore important for local species persistence under environmental change (Bell and Gonzalez 2009). This intraspecific genetic variation represents the potential for further adaptive change in response to new selective challenges such as the global warming (Rice and Emery 2003). Thus, understanding the adaptive genetic variation related to the hot drought may help us better predict and manage forests in a changing climate (Neale and Kremer 2011, Oney et al. 2013).

Many researchers have investigated drought-tolerant physiological traits of conifers trees (Teskey et al. 1987, Cregg and Zhang 2001, McDowell et al. 2008, McDowell 2011). Multiple traits can affect the drought responses of conifer trees, such as root-to-shoot ratio, root biomass and length, specific leaf area (SLA, the ratio of leaf area to dry mass), stomatal conductance, and water use efficiency (WUE, ratio between CO_2 assimilation and transpiration; Picon et al. 1996, Cregg and Zhang 2001, de Miguel et al. 2012, Olmo et al. 2014, Moran et al. 2017). Stomatal regulation and structural adjustments in leaf area minimize water loss, while deep root systems maximize the potential for water uptake. However, which traits are most important for drought tolerance in which species or circumstances remains largely unknown.

Most recent research on drought stress has focused on aboveground tree parts (McDowell et al. 2008, Ryan 2011, Hamanishi and Campbell 2011), while belowground traits are largely missing due to the difficulties in observing and studying roots (McDowell et al. 2008, Hamanishi and Campbell 2011, Brunner et al. 2015). For adult trees in the field, retrieval of all roots, or measurement of their maximum length, is a challenge (Robinson 2004). Tree roots are not only responsible for water uptake, but also act as sensors for water-deficit conditions, sending signals to shoots (Brunner and Godbold 2007, Hamanishi and Campbell 2011). Studies have shown the critical role of roots in drought responses of both adult trees and seedlings. For example, under severe drought conditions, studies showed that trees in the field tend to increase root-to-shoot ratios and root biomass at the expense of stems (Mokany et al. 2006, Poorter et al. 2012). Seedlings of trees also increase allocation of biomass to roots to augment water acquisition during drought (Markesteijn and Poorter 2009). This allocation allows the plants to increase water uptake, while reducing potential water loss from transpiration that can occur when more leaves are produced. However, extended drought can lead to root die-back, and reduced capability for water uptake (Eldhuset et al. 2013, Plaut et al. 2013). Thus, incorporating the traits of roots is necessary to understand the drought tolerance of forest trees.

Multiple provenance studies have identified patterns consistent with local adaptation to drought (Moran et al. 2017). For example, trees from drier climates often exhibit slower height or needle growth (de la Mata et al. 2014), less aboveground biomass or a shorter growing season (Kerr et al. 2015). Moreover, seedlings from dry environments often exhibit more root growth, higher drought survival (Cregg and Zhang 2001, Kolb et al. 2016), and higher WUE (Cregg et al. 2000, Voltas et al. 2008). Genetic differences are likely to play an important role in geographical variation in these drought-tolerance traits (McDowell et al. 2008). However, the genes underlying these drought-tolerance traits are mostly unknown. Several studies have investigated changes in gene expression in drought-stressed conifer seedlings (e.g., Ralph et al. 2006, Hamanishi and

Campbell 2011). Some genes may relate to drought tolerance, such as those involved in late-embryogenesis-abundant (LEA) proteins, abscissic acid (ABA) signaling pathways, and carbohydrate and lipid metabolism (Ralph et al. 2006, Hamanishi and Campbell 2011). However, most of the gene expression changes return to normal after re-watering the drought-stressed seedlings. Such changes are responsible for plastic environmental responses, rather than locally adaptive differences in mean traits (Bräutigam et al. 2013). Some fundamental questions are still largely unresolved, including the nature and number of genes involved in adaptation to drought (Barton and Keightley 2002, Prunier et al. 2011).

1.3 Focal species

I chose *P. ponderosa* as the focal species for four reasons. First, it is a major source of timber that covers 27 million acres in western North America (Schubert 1974, Oliver and Ryker 1990), and provides important ecosystem services (e.g., habitat for wildlife) (Burns and Honkala 1990). Second, this species tolerates a wide range of temperatures and precipitation (Conkle and Critchfield 1988), and is regarded as one of the most drought-tolerant trees in North America (Kolb and Robberecht 1996). Third, genetic resources are fairly well developed in conifers compared to other woody plants (Pavy et al. 2016), meaning that this study can use the available methods of sequencing and data analysis, as well as the available reference genome of loblolly pine (Zimin et al. 2014). Despite the importance of *P. ponderosa*, no previous study has investigated the genetic basis of drought tolerance in this species.

In the 1970s, the Forest Service's Pacific Southwest Regional Genetic Resources Program planted clones of 302 wild ponderosa pine genotypes in Chico, California. They came from diverse climate conditions in the central portion of California's Sierra Nevada mountains and are now reproductively mature, thus presenting an excellent resource for genetic studies (Figure 1.1). For this study, I chose 223 individual *P. ponderosa* genotypes from the orchard collection. The source locations of these 223 genotypes are shown in Figure 1.2. These locations likely fall within just one of the several genetic subdivisions previously identified in ponderosa pine (Conkle and Critchfield 1988, Willyard et al. 2009, Potter et al. 2015).

1.4 Genetic association techniques

Recently, there has been an increase in the use of genetic association techniques to identify genes underlying quantitative traits in forest trees (Eckert et al. 2010a, Riordan et al. 2016, Di Pierro et al. 2017). A quantitative trait is a phenotype that exhibits continuous variation due to the cumulative actions of many genes. In an association analysis, regression is used to identify variable genetic markers statistically associated with either a phenotype of interest or an environment. In species with large genomes and/or low linkage disequilibrium like conifers, large number of markers distributed over the genome are required in order to identify all or most of the important genes (Schlötterer 2004). Single nucleotide polymorphisms (SNPs), meaning sites where individual sequences differ by a single base pair, are the most abundant type of polymorphism and are often used as genetic markers in association studies (Schlötterer 2004). With the dropping cost of sequencing, approaches that generate thousands of

SNPs are increasingly being used. For instance, Genotyping-by-Sequencing (GBS), which involves the use of restriction enzymes to cut and sequence a small subset of the genome (Elshire et al. 2011, Andrews et al. 2016), can produce tens of thousands of SNPs with high coverage (Chen et al. 2013a, Pan et al. 2015). In addition, GBS can genotype tree species with or without the availability of a reference genome (Chen et al. 2013a). Thus, GBS is an efficient and affordable approach to obtain SNP data.

Uncovering the genetic basis of adaptation hinges on the ability to detect loci under selection. There are two types of genetic association techniques. Genotype-to-environment (G2E) association investigates the statistical association between genetic variation at individual loci and the environment, which can be useful in the search for genes responsible for local adaptation (Eckert et al. 2010b, 2010a). For example, an association between a SNP and aridity may indicate that the gene or its regulatory region affects trees performance in dry versus wet environments. However, G2E studies do not reveal how SNPs are connected to phenotypic differences, and thus what traits are under selection in a given environment. This is where Genotype-to-Phenotype (G2P) association studies are useful. G2P association identifies loci linked to a particular phenotype by correlating genotypes at SNPs with the variation in certain traits (Eckert et al. 2009, Holliday et al. 2010). To eliminate the effects of environment on phenotypes, traits must be measured in a common environment, such as a greenhouse. However, G2P association study does not reveal how or if the trait affects the fitness of trees in the field. G2E and G2P association are thus complementary (Figure 1.3). Although the combination of these two approaches can help to identify genes and traits under selection in natural settings, very few studies have combined them (Eckert et al. 2009, 2015, Moran et al. 2017). By combining G2E and G2P association genetics, this study can identify both the loci and traits under selection of ponderosa pine in nature.

1.5 Objectives of this study

The main purpose of this study is to understand the adaptive genetic variation as well as the genetic basis of adaptive phenotypes of a non-model tree species, *P. ponderosa*. Three types of data were necessary to conduct this research. First, I obtained the raw genetic data from GBS. Second, I obtained a 30-year (1951-1980) averages of climate data from the 270 m resolution California Basin Characterization Model (BCM) (Flint et al. 2013). Third, I obtained the phenotypic data by conducting a greenhouse experiment.

In Chapter 2, I present a comparison of multiple pipelines for SNP calling in pine. Genotyping and identifying SNPs is a challenge in most conifer species due to their extremely large (19-32 Gb) and highly repetitive genomes (Birol et al. 2013, Zimin et al. 2014, Neale et al. 2014, Stevens et al. 2016). Before I ran GBS, it was beneficial to determine which enzyme produces the most fragments within the desired size range (100-400 bp). For optimization of the GBS protocol, 1000 ng samples of *P. ponderosa* genomic DNA were digested separately with *Ape*KI, *Pst*I, and *Eco*T22I. *Ape*KI yielded a high smooth curve of fragment sizes between 150 and 500, which indicates good performance for GBS. GBS has been tested for conifers on small numbers of individuals (<10) and has been found produce tens of thousands of SNPs with high coverage (Chen et al. 2013b, Pan et al. 2015). However, the use of GBS on conifers species is still largely

limited by the difficulty of genome-wide SNP calling from the massively parallel short-read sequences (Glenn 2011, Goto et al. 2017). Even though GBS only sequences a fraction of the genome, conifer genomes are so large and repetitive the datasets produced still present a computational challenge. I compared four GBS bioinformatics pipelines, two of which require a reference genome (TASSEL-GBS V2 and Stacks), two of which are *de novo* pipelines (UNEAK and Stacks). I used Illumina sequence data from 94 ponderosa pines, with loblolly pine as the reference genome. The number of SNPs called was much lower without a reference genome (62 -196 thousand vs. 2.1 - 2.7 million SNPs). UNEAK was the fastest overall and identified more SNPs than Stacks *de novo*. Stacks with a reference genome produced the highest number of SNPs with lowest proportion of paralogs, while SNPs identified by TASSEL-GBS V2 exhibited the highest heterozygosity, minor allele frequency, and proportion of paralogs. More SNPs were uniquely identified by Stacks than TASSEL, though there was high overlap between methods. Researchers studying other conifer species should be prepared to analyze very large numbers of SNPs, and to consider the benefits and limitations of different pipelines.

Chapter 3 focuses on the adaptive genetic variation of *P. ponderosa* by running G2E association analysis using the SNPs produced from the Stacks reference-based pipeline, as well as gene annotation of the significantly environmentally associated SNPs. The Stacks reference-based pipeline identified 4,155,896 SNPs (Chapter 2). I selected five variables for G2E analysis with generally low-to-moderate correlations between them (Figure 1.4). These five variables include: mean climatic water deficit (CWD, a measure of evaporative demand exceeding soil moisture); mean minimum temperature of winter (TMIN), calculated as the average minimum temperature over the coldest months (December–February); mean maximum temperature of summer (TMAX), calculated as the average maximum temperature over the hottest months (June–August); mean monthly precipitation of winter (PPTW), calculated as the average monthly precipitation over the coldest months (December–February); and April 1st snow pack (PCK4). There are several models available for G2E, such as BAYENV (Günther and Coop 2013), BAYPASS (Gautier 2015), BAYESCENV (Villemereuil and Gaggiotti 2015), and latent factor mixed model (LFMM; Frichot et al. 2013, Frichot and François 2015). These models can effectively account for population structure, but can be computationally intensive and slow with Markov chain Monte Carlo algorithms or Bayesian bootstrap methods. To conduct G2E analysis with the 4,155,896 SNPs and five chosen climatic variables, we chose LFMM2, which was developed for G2E association and has been shown to outperform other similar approaches with several orders-of-magnitude faster computing (Caye et al. 2019). I found 213 SNPs strongly associated with mean maximum temperature of summer (TMAX), 335 with mean minimum temperature of winter (TMIN), 1798 with April 1st snow pack (PCK4), and 120 SNPs with mean climatic water deficit (CWD), but no SNP with mean monthly precipitation of winter (PPTW). Different protein functions have been annotated underlying the genetic associations, including ubiquitination, abscisic acid (ABA) signaling pathway, cell division or growth of roots or shoots, cell wall organization, and seed dormancy.

Chapter 4 focuses on the adaptive phenotype of *P. ponderosa* by running the G2P association analysis as well as gene annotation of the significantly phenotypically associated SNPs. To obtain the phenotypic data, I conducted a greenhouse experiment

with seeds from 48 already genotyped mother trees. In the greenhouse experiment, 10 seedlings for each mother trees were planted in both dry and wet treatment. Eight phenotypic traits were measured during or after the greenhouse experiment. Based on ANOVA analysis with these phenotypic data from wet and dry treatments while accounting for block (planting box-level) differences, six drought responsive traits were identified, including: treatment, RL (root length), SW (shoot weight), R2S (root-shoot dry mass ratio), SDAD (stomatal density on adaxial side), NRAB (number of stomatal rows on abaxial side), and GR (growth). In the drought treatment, seedlings present larger RL, R2S, SDAD, NRAB, and lower GR and SW. Then, I ran a G2P analysis with previous genotype (chapter 3) and these 6 traits using LFMM2. I found 153 SNPs strongly associated with RL, 80 with SW, 145 with GR, 42 with SDAD, 85 with NRAB, and 1530 with R2S. The identified SNPs reside in genes with a wide variety of functions, including ubiquitination, abscisic acid (ABA) signaling pathway, cell division or growth of roots or shoots, and cell wall organization. Potentially, the identified genes and alleles are valuable resources for pine trees breeding through marker assisted selection and genomic selection, specifically under the rapid changing climate scenarios. In addition, roots play a critical role in both the drought responsive traits and the function of correlated genes in our study. Future studies may need to incorporate the root traits to understand the response of pine trees to changing climate.

1.6 Significance

Understanding the genetic basis of local adaptation to climate in the context of global change poses one of the greatest challenges of this century (Manel et al. 2010). Some studies have begun to explore the molecular basis of phenotypic traits associated with local adaptation in model species with available whole genome sequences (Umina et al. 2005, Begun et al. 2007). However, for most non-model species, the genetic basis of adaptation is unknown (Eckert et al. 2009).

This dissertation evaluated how different traits contribute to overall drought tolerance in seedlings, identified genetic loci associated with the measured traits and individual growth and survival, and identified what genes were responsible for local adaptation. This dissertation focused on multiple drought response traits, several climate variables, and a large number of SNPs, which can help to identify genes that have variation associated with both environmental gradients, drought tolerance traits, or both. This work has direct implications for forest management and conservation, such as identifying seed sources with drought-tolerance related genes for restoration and plantations. Such seedlings may have a higher survival and growth rate compared to other seed sources (Goodrich et al. 2016). Also, this association genetics study in ponderosa pine greatly enhances our understanding of the adaptation of other conifers and plants under changing climate. In addition, this study provides clues for what genes or traits might be important for drought tolerance in other species.

1.7 References

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. Evolutionary Applications 1:95–111.

Alberto, F. J., S. N. Aitken, R. Alía, S. C. González-Martínez, H. Hänninen, A. Kremer, F. Lefèvre, T. Lenormand, S. Yeaman, R. Whetten, and O. Savolainen. 2013. Potential for evolutionary responses to climate change – evidence from tree populations. Global Change Biology 19:1645–1661.

Allen, C. D., A. K. Macalady, H. Chenchouni, D. Bachelet, N. McDowell, M. Vennetier, T. Kitzberger, A. Rigling, D. D. Breshears, E. H. (Ted) Hogg, P. Gonzalez, R. Fensham, Z. Zhang, J. Castro, N. Demidova, J.-H. Lim, G. Allard, S. W. Running, A. Semerci, and N. Cobb. 2010. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. Forest Ecology and Management 259:660–684.

Anderson, J. T., A. M. Panetta, and T. Mitchell-Olds. 2012. Evolutionary and Ecological Responses to Anthropogenic Climate Change: Update on Anthropogenic Climate Change. Plant Physiology 160:1728–1740.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature reviews. Genetics 17:81–92.

Barton, N. H., and P. D. Keightley. 2002. Understanding quantitative genetic variation. Nature Reviews Genetics 3:11–21.

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y.-P. Poh, M. W. Hahn, P. M. Nista, C. D. Jones, A. D. Kern, C. N. Dewey, L. Pachter, E. Myers, and C. H. Langley. 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. PLOS Biology 5:e310.

Bell, G., and A. Gonzalez. 2009. Evolutionary rescue can prevent extinction following environmental change. Ecology Letters 12:942–948.

Berg, N., and A. Hall. 2015. Increased Interannual Precipitation Extremes over California under Climate Change. Journal of Climate 28:6324–6334.

Birol, I., A. Raymond, S. D. Jackman, S. Pleasance, R. Coope, G. A. Taylor, M. M. S. Yuen, C. I. Keeling, D. Brand, B. P. Vandervalk, H. Kirk, P. Pandoh, R. A. Moore, Y. Zhao, A. J. Mungall, B. Jaquish, A. Yanchuk, C. Ritland, B. Boyle, J. Bousquet, K. Ritland, J. MacKay, J. Bohlmann, and S. J. M. Jones. 2013. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497.

Bräutigam, K., K. J. Vining, C. Lafon-Placette, C. G. Fossdal, M. Mirouze, J. G. Marcos, S. Fluch, M. F. Fraga, M. Á. Guevara, D. Abarca, Ø. Johnsen, S. Maury, S. H. Strauss, M. M. Campbell, A. Rohde, C. Díaz-Sala, and M.-T. Cervera. 2013. Epigenetic regulation of adaptive responses of forest tree species to the environment. Ecology and Evolution 3:399–415.

Brunner, I., and D. L. Godbold. 2007. Tree roots in a changing world. Journal of Forest Research 12:78–82.

Brunner, I., C. Herzog, M. A. Dawes, M. Arend, and C. Sperisen. 2015. How tree roots respond to drought. Frontiers in Plant Science 6.

Burns, R. M., and B. H. (Technical C. Honkala. 1990. Silvics of North America. Volume 1. Conifers. Agriculture Handbook (Washington).

Caye, K., B. Jumentier, J. Lepeule, and O. François. 2019. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Molecular Biology and Evolution 36:852–860.

Chen, C., S. E. Mitchell, R. J. Elshire, E. S. Buckler, and Y. A. El-Kassaby. 2013a. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genetics & Genomes 9:1537–1544.

Chen, C., S. E. Mitchell, R. J. Elshire, E. S. Buckler, and Y. A. El-Kassaby. 2013b. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genetics & Genomes 9:1537–1544.

Conkle, M. T., and W. B. Critchfield. 1988. Genetic variation and hybridization of ponderosa pine. In: Ponderosa Pine: the species and its management, Washington State University Cooperative Extension, 1988: p. 27-43.

Cregg, B. M., J. M. Olivas-García, and T. C. Hennessey. 2000. Provenance variation in carbon isotope discrimination of mature ponderosa pine trees at two locations in the Great Plains. Canadian Journal of Forest Research 30:428–439.

Cregg, B. M., and J. W. Zhang. 2001. Physiology and morphology of Pinus sylvestris seedlings from diverse sources under cyclic drought stress. Forest Ecology and Management 154:131–139.

Di Pierro, E. A., E. Mosca, S. C. González-Martínez, G. Binelli, D. B. Neale, and N. La Porta. 2017. Adaptive variation in natural Alpine populations of Norway spruce (Picea abies [L.] Karst) at regional scale: Landscape features and altitudinal gradient effects. Forest Ecology and Management 405:350–359.

Eckert, A. J., A. D. Bower, S. C. GonzáLez-MartíNez, J. L. Wegrzyn, G. Coop, and D. B. Neale. 2010a. Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). Molecular Ecology 19:3789–3805.

Eckert, A. J., A. D. Bower, J. L. Wegrzyn, B. Pande, K. D. Jermstad, K. V. Krutovsky, J. B. St. Clair, and D. B. Neale. 2009. Association Genetics of Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-Hardiness Related Traits. Genetics 182:1289–1302.

Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra, S. C. Gonzalez-Martinez, and D. B. Neale. 2010b. 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., P. E. Maloney, D. R. Vogler, C. E. Jensen, A. D. Mix, and D. B. Neale. 2015. Local adaptation at fine spatial scales: an example from sugar pine (Pinus lambertiana, Pinaceae). Tree Genetics & Genomes 11.

Eilmann, B., and A. Rigling. 2012. Tree-growth analyses to estimate tree species' drought tolerance. Tree Physiology 32:178–187.

Eldhuset, T. D., N. E. Nagy, D. Volařík, I. Børja, R. Gebauer, I. A. Yakovlev, and P. Krokene. 2013. Drought affects tracheid structure, dehydrin expression, and above- and belowground growth in 5-year-old Norway spruce. Plant and Soil 366:305–320.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6:e19379.

Fettig, C. J., L. A. Mortenson, B. M. Bulaon, and P. B. Foulk. 2019. Tree mortality following drought in the central and southern Sierra Nevada, California, U.S. Forest Ecology and Management 432:164–178.

Flint, L. E., A. L. Flint, J. H. Thorne, and R. Boynton. 2013. Fine-scale hydrologic modeling for regional landscape applications: the California Basin Characterization Model development and performance. Ecological Processes 2:1–21.

Frichot, E., and O. François. 2015. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6:925–929.

Frichot, E., S. D. Schoville, G. Bouchard, and O. François. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular Biology and Evolution 30:1687–1699.

Gautier, M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics 201:1555–1579.

Giorgi, F., and P. Lionello. 2008. Climate change projections for the Mediterranean region. Global and Planetary Change 63:90–104.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11:759–769.

Goodrich, B. A., K. M. Waring, and T. E. Kolb. 2016. Genetic variation in Pinus strobiformis growth and drought tolerance from southwestern US populations. Tree Physiology 36:1219–1235.

Goto, S., H. Kajiya-Kanegae, W. Ishizuka, K. Kitamura, S. Ueno, Y. Hisamoto, H. Kudoh, M. Yasugi, A. J. Nagano, and H. Iwata. 2017. Genetic mapping of local adaptation along the altitudinal gradient in Abies sachalinensis. Tree Genetics & Genomes 13:104.

Goulden, M. L., and R. C. Bales. 2014. Mountain runoff vulnerability to increased evapotranspiration with vegetation expansion. Proceedings of the National Academy of Sciences 111:14071–14075.

Griffin, D., and K. J. Anchukaitis. 2014. How unusual is the 2012–2014 California drought? Geophysical Research Letters 41:2014GL062433.

Günther, T., and G. Coop. 2013. Robust Identification of Local Adaptation from Allele Frequencies. Genetics 195:205–220.

Hamanishi, E. T., and M. M. Campbell. 2011. Genome-wide responses to drought in forest trees. Forestry: An International Journal of Forest Research 84:273–283.

Holliday, J. A., K. Ritland, and S. N. Aitken. 2010. Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (Picea sitchensis). New Phytologist 188:501–514.

IPCC. 2014. Climate Change 2014: Impacts,Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Page 1132. Cambridge University Press, Cambridge, UK and New York, USA.

J, M. 1977. California climate in relation to vegetation.

Kerr, K., F. Meinzer, K. McCulloh, D. Woodruff, and D. Marias. 2015. Expression of functional traits during seedling establishment in two populations of *Pinus ponderosa* from contrasting climates. Tree Physiology 35:535–548.

Kitzmiller, J. H. 2005. Provenance Trials of Ponderosa Pine in Northern California. Forest Science 51:595–607.

Kolb, P. F., and R. Robberecht. 1996. High temperature and drought stress effects on survival of Pinus ponderosa seedlings. Tree Physiology 16:665–672.

Kolb, T. E., K. C. Grady, M. P. McEttrick, and A. Herrero. 2016. Local-Scale Drought Adaptation of Ponderosa Pine Seedlings at Habitat Ecotones. Forest Science 62:641–651.

Langlet, O. 1971. Two Hundred Years Genecology. Taxon 20:653–721.

Lloret, F., E. G. Keeling, and A. Sala. 2011. Components of tree resilience: effects of successive low-growth episodes in old ponderosa pine forests. Oikos 120:1909–1920.

Loarie, S. R., P. B. Duffy, H. Hamilton, G. P. Asner, C. B. Field, and D. D. Ackerly. 2009. The velocity of climate change. Nature 462:1052–1055.

Manel, S., S. Joost, B. K. Epperson, R. Holderegger, A. Storfer, M. S. Rosenberg, K. T. Scribner, A. Bonin, and M.-J. Fortin. 2010. Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. Molecular Ecology 19:3760–3772.

van Mantgem, P. J., N. L. Stephenson, J. C. Byrne, L. D. Daniels, J. F. Franklin, P. Z. Fule, M. E. Harmon, A. J. Larson, J. M. Smith, A. H. Taylor, and T. T. Veblen. 2009. Widespread Increase of Tree Mortality Rates in the Western United States. Science 323:521–524.

Markesteijn, L., and L. Poorter. 2009. Seedling root morphology and biomass allocation of 62 tropical tree species in relation to drought- and shade-tolerance. Journal of Ecology 97:311–325.

de la Mata, R., E. Merlo, and R. Zas. 2014. Among-population variation and plasticity to drought of Atlantic, Mediterranean, and interprovenance hybrid populations of maritime pine. Tree Genetics & Genomes 10:1191–1203.

McDowell, N. G. 2011. Mechanisms Linking Drought, Hydraulics, Carbon Metabolism, and Vegetation Mortality. Plant Physiology 155:1051–1059.

McDowell, N. G., A. P. Williams, C. Xu, W. T. Pockman, L. T. Dickman, S. Sevanto, R. Pangle, J. Limousin, J. Plaut, D. S. Mackay, J. Ogee, J. C. Domec, C. D. Allen, R. A. Fisher, X. Jiang, J. D. Muss, D. D. Breshears, S. A. Rauscher, and C. Koven. 2015. Multi-scale predictions of massive conifer mortality due to chronic temperature rise. Nature Climate Change 6:295–300.

McDowell, N., W. T. Pockman, C. D. Allen, D. D. Breshears, N. Cobb, T. Kolb, J. Plaut, J. Sperry, A. West, D. G. Williams, and E. A. Yepez. 2008. Mechanisms of plant survival and mortality during drought: why do some plants survive while others succumb to drought? New Phytologist 178:719–739.

de Miguel, M., D. Sanchez-Gomez, M. T. Cervera, and I. Aranda. 2012. Functional and genetic characterization of gas exchange and intrinsic water use efficiency in a full-sib family of *Pinus pinaster* Ait. in response to drought. Tree Physiology 32:94–103.

Mokany, K., R. J. Raison, and A. S. Prokushkin. 2006. Critical analysis of root : shoot ratios in terrestrial biomes. Global Change Biology 12:84–96.

Montwé, D., M. Isaac-Renton, A. Hamann, and H. Spiecker. 2016. Drought tolerance and growth in populations of a wide-ranging tree species indicate climate change risks for the boreal north. Global Change Biology 22:806–815.

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

Neale, D. B., and A. Kremer. 2011. Forest tree genomics: growing resources and applications. Nature Reviews Genetics 12:111–122.

Neale, D. B., J. L. Wegrzyn, K. A. Stevens, A. V. Zimin, D. Puiu, M. W. Crepeau, C. Cardeno, M. Koriabine, A. E. Holtz-Morris, J. D. Liechty, P. J. Martínez-García, H. A. Vasquez-Gross, B. Y. Lin, J. J. Zieve, W. M. Dougherty, S. Fuentes-Soriano, L.-S. Wu, D. Gilbert, G. Marçais, M. Roberts, C. Holt, M. Yandell, J. M. Davis, K. E. Smith, J. F. Dean, W. W. Lorenz, R. W. Whetten, R. Sederoff, N. Wheeler, P. E. McGuire, D. Main, C. A. Loopstra, K. Mockaitis, P. J. deJong, J. A. Yorke, S. L. Salzberg, and C. H. Langley. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology 15:R59.

Newton, R. J., E. A. Funkhouser, F. Fong, and C. G. Tauer. 1991. Molecular and physiological genetics of drought tolerance in forest species. Forest Ecology and Management 43:225–250.

Oliver, W. W., and R. A. Ryker. 1990. Pinus ponderosa Dougl. ex Laws. ponderosa pine. Silvics of North America 1:413.

Olmo, M., B. Lopez-Iglesias, and R. Villar. 2014. Drought changes the structure and elemental composition of very fine roots in seedlings of ten woody tree species. Implications for a drier climate. Plant and Soil 384:113–129.

Oney, B., B. Reineking, G. O'Neill, and J. Kreyling. 2013. Intraspecific variation buffers projected climate change impacts on Pinus contorta. Ecology and Evolution 3:437–449.

Pachauri, R. K., M. R. Allen, V. R. Barros, J. Broome, W. Cramer, R. Christ, J. A. Church, L. Clarke, Q. Dahe, and P. Dasgupta. 2014. Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. IPCC.

Pan, J., B. Wang, Z.-Y. Pei, W. Zhao, J. Gao, J.-F. Mao, and X.-R. Wang. 2015. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Molecular Ecology Resources 15:711–722.

Pavy, N., F. Gagnon, A. Deschênes, B. Boyle, J. Beaulieu, and J. Bousquet. 2016. Development of highly reliable in silico SNP resource and genotyping assay from exome capture and sequencing: an example from black spruce (*Picea mariana*). Molecular Ecology Resources 16:588–598.

Pereira, H. M., P. W. Leadley, V. Proença, R. Alkemade, J. P. Scharlemann, J. F. Fernandez-Manjarrés, M. B. Araújo, P. Balvanera, R. Biggs, and W. W. Cheung. 2010. Scenarios for global biodiversity in the 21st century. Science 330:1496–1501.

Picon, C., J. M. Guehl, and A. Ferhi. 1996. Leaf gas exchange and carbon isotope composition responses to drought in a drought-avoiding (Pinus pinaster) and a drought-tolerant (Quercus petraea) species under present and elevated atmospheric CO2 concentrations. Plant, Cell & Environment 19:182–190.

Plaut, J. A., W. D. Wadsworth, R. Pangle, E. A. Yepez, N. G. McDowell, and W. T. Pockman. 2013. Reduced transpiration response to precipitation pulses precedes mortality in a piñon–juniper woodland subject to prolonged drought. New Phytologist 200:375–387.

Poorter, H., K. J. Niklas, P. B. Reich, J. Oleksyn, P. Poot, and L. Mommer. 2012. Biomass allocation to leaves, stems and roots: meta-analyses of interspecific variation and environmental control. New Phytologist 193:30–50.

Potter, K. M., V. D. Hipkins, M. F. Mahalovich, and R. E. Means. 2015. Nuclear genetic variation across the range of ponderosa pine (Pinus ponderosa): Phylogeographic, taxonomic and conservation implications. Tree Genetics & Genomes 11:38.

Prunier, J., J. Laroche, J. Beaulieu, and J. Bousquet. 2011. Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce: SNPs and climate adaptation. Molecular Ecology 20:1702–1716.

Ralph, S. G., H. Yueh, M. Friedmann, D. Aeschliman, J. A. Zeznik, C. C. Nelson, Y. S. Butterfield, R. Kirkpatrick, J. Liu, and S. J. Jones. 2006. Conifer defence against insects: microarray gene expression profiling of Sitka spruce (Picea sitchensis) induced by mechanical wounding or feeding by spruce budworms (Choristoneura occidentalis) or white pine weevils (Pissodes strobi) reveals large-scale changes of the host transcriptome. Plant, Cell & Environment 29:1545–1570.

Rice, K. J., and N. C. Emery. 2003. Managing microevolution: restoration in the face of global change. Frontiers in Ecology and the Environment 1:469–478.

Riordan, E. C., P. F. Gugger, J. Ortego, C. Smith, K. Gaddis, P. Thompson, and V. L. Sork. 2016. Association of genetic and phenotypic variability with geography and climate in three southern California oaks. American Journal of Botany 103:73–85.

Robinson, D. 2004. Scaling the depths: below-ground allocation in plants, forests and biomes. Functional Ecology 18:290–295.

Royce, E. B., and M. G. Barbour. 2001. Mediterranean climate effects. I. Conifer water use across a Sierra Nevada ecotone. American Journal of Botany 88:911–918.

Ryan, M. G. 2011. Tree responses to drought. Tree Physiology 31:237–239.

Schlötterer, C. 2004. The evolution of molecular markers — just a matter of fashion? Nature Reviews Genetics 5:63–69.

Schubert, G. H. 1974. Silviculture of southwestern ponderosa pine: The status of our knowledge. Res. Pap. RM-123. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station. 71 p. 123.

Schultz, G. 2017. Tree Mortality: Facts and Figures:19.

Stevens, K. A., J. L. Wegrzyn, A. Zimin, D. Puiu, M. Crepeau, C. Cardeno, R. Paul, D. Gonzalez-Ibeas, M. Koriabine, A. E. Holtz-Morris, P. J. Martínez-García, U. U. Sezen, G. Marçais, K. Jermstad, P. E. McGuire, C. A. Loopstra, J. M. Davis, A. Eckert, P. de Jong, J. A. Yorke, S. L. Salzberg, D. B. Neale, and C. H. Langley. 2016. Sequence of the Sugar Pine Megagenome. Genetics 204:1613–1626.

Teskey, R. O., B. C. Bongarten, B. M. Cregg, P. M. Dougherty, and T. C. Hennessey. 1987. Physiology and genetics of tree growth response to moisture and temperature stress: an examination of the characteristics of loblolly pine (Pinus taeda L.). Tree Physiology 3:41–61.

Umina, P. A., A. R. Weeks, M. R. Kearney, S. W. McKechnie, and A. A. Hoffmann. 2005. A Rapid Shift in a Classic Clinal Pattern in Drosophila Reflecting Climate Change. Science 308:691–693.

Villemereuil, P. de, and O. E. Gaggiotti. 2015. A new FST-based method to uncover local adaptation using environmental variables. Methods in Ecology and Evolution 6:1248–1258.

Voltas, J., M. R. Chambel, M. A. Prada, and J. P. Ferrio. 2008. Climate-related variability in carbon and oxygen stable isotopes among populations of Aleppo pine grown in common-garden tests. Trees 22:759–769.

Williams, A. P., R. Seager, J. T. Abatzoglou, B. I. Cook, J. E. Smerdon, and E. R. Cook. 2015. Contribution of anthropogenic warming to California drought during 2012–2014. Geophysical Research Letters 42:2015GL064924.

Willyard, A., R. Cronn, and A. Liston. 2009. Reticulate evolution and incomplete lineage sorting among the ponderosa pines. Molecular Phylogenetics and Evolution 52:498–511.

Wright, J. W. 2007. Local adaptation to serpentine soils in Pinus ponderosa. Plant and Soil 293:209–217.

Ying, C. C., and Q. Liang. 1994. Geographic pattern of adaptive variation of lodgepole pine (Pinus contorta Dougl.) within the species' coastal range: field performance at age 20 years. Forest Ecology and Management 67:281–298.

Zimin, A., K. A. Stevens, M. W. Crepeau, A. Holtz-Morris, M. Koriabine, G. Marçais, D. Puiu, M. Roberts, J. L. Wegrzyn, P. J. de Jong, D. B. Neale, S. L. Salzberg, J. A. Yorke, and C. H. Langley. 2014. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics 196:875–890.

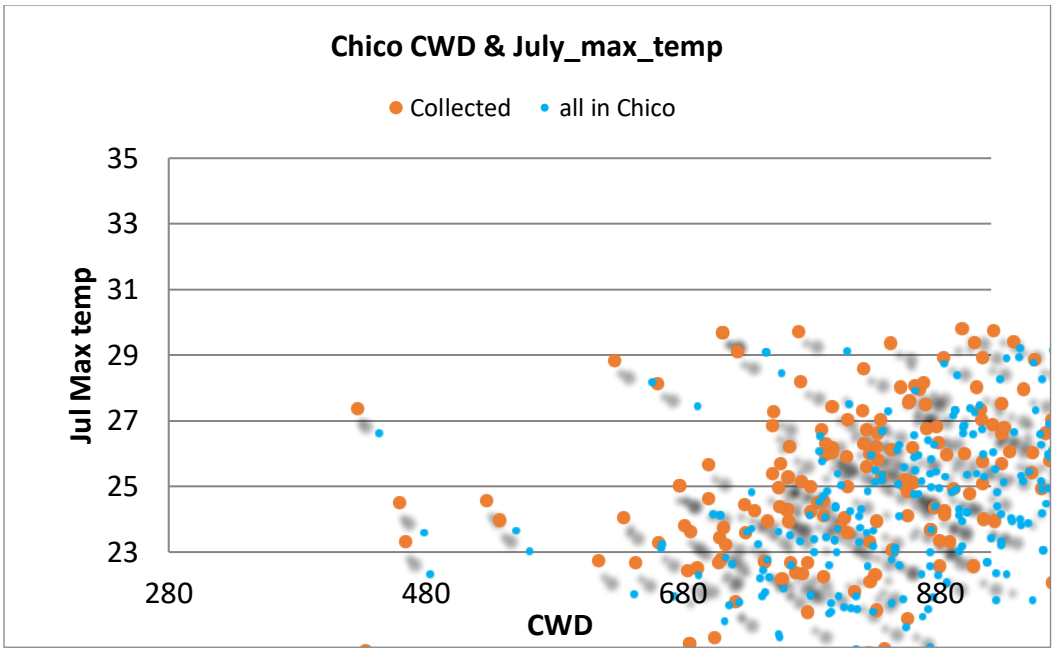

Figure 1.1 CWD (mm) and July maximum temperature (°C) for 302 *P. ponderosa* in the Chico orchard. The orange dots represent the 223 genotypes from which we collected needles. The blue dots represent all the 302 genotypes in the Chico orchard

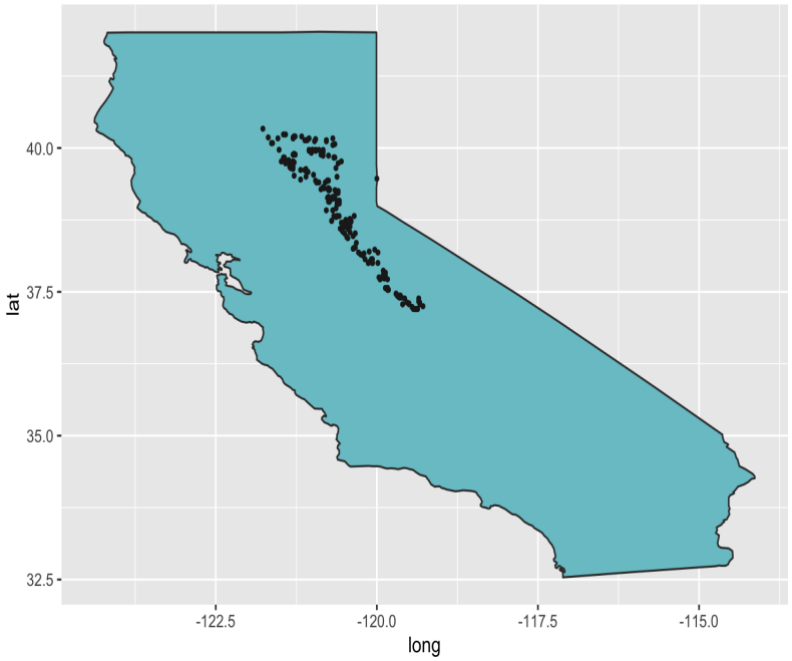

Figure 1.2 Geographic distribution of the 223 samples. The black dots represent original genotype source locations.

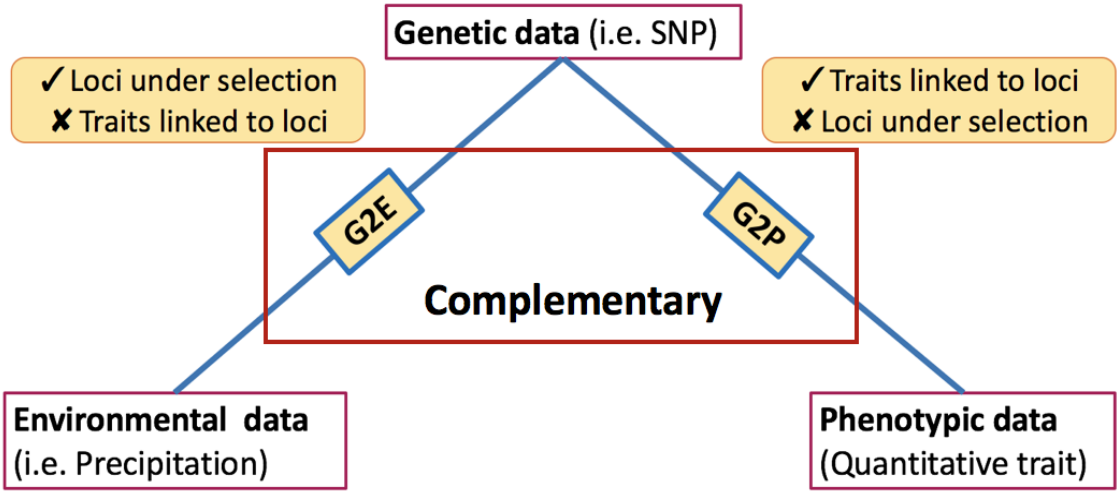

Figure 1.3 Genetic association techniques: type of data involved and types of association. G2E represents genotype-to-environment association. G2P represents genotype-to-phenotype association.

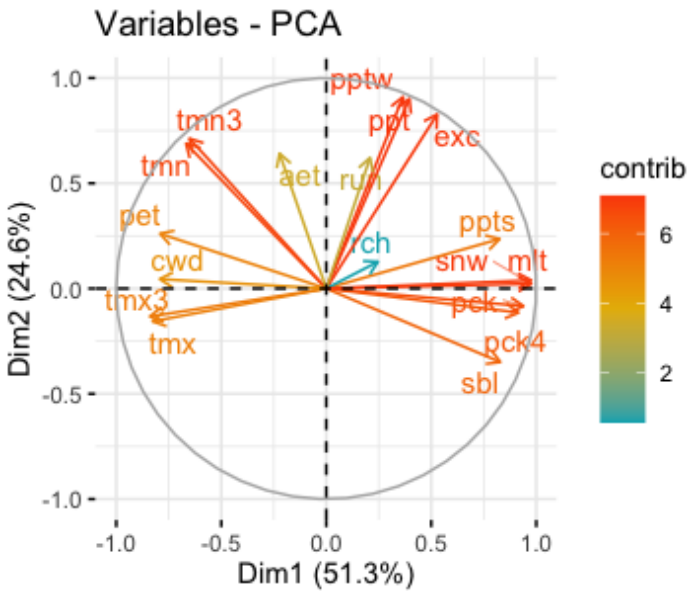

Figure 1.4 PCA analysis of 30-year averages (1951-1980) of all the 18 environmental variables from BCM model.

Chapter 2:
Testing pipelines for genome-wide SNP calling from Genotyping-By-Sequencing (GBS) data for *Pinus ponderosa*

2.0 Abstract

Single Nucleotide Polymorphism (SNP) markers have rapidly gained popularity due to their abundance in most genomes and their amenability to high-throughput genotyping techniques. Genotyping-by-sequencing (GBS) has been demonstrated to be a robust and cost-effective genotyping method. While previous studies have shown that alignment of the short-read fragments to a genome sequence results in better SNP calling than *de novo* approaches, only a few tree species - and few conifers in particular - have an annotated sequence. While these could be used to align sequence fragments from related species, sequence divergence might result in SNPs being missed if they are in fragments that do not align properly. Producing a new annotated genome sequence for every conifer species before SNP analyses are conducted is still prohibitive, as many conifer genomes are huge (>19 GB) and include a large proportion of repeat sequences, making assembly difficult. Here we compare four GBS bioinformatics pipelines, and two of which require a reference genome (TASSEL-GBS V2 and Stacks), two of which are *de novo* pipelines (UNEAK and Stacks). We used Illumina sequence data from 94 ponderosa pines, with loblolly pine as the reference genome. The number of SNPs called was much lower without a reference genome (62 -196 thousand vs. 2.1 - 2.7 million SNPs). UNEAK was the fastest overall and identified more SNPs than Stacks *de novo*. Stacks with a reference genome produced the highest number of SNPs with lowest proportion of paralogs, while SNPs identified by TASSEL-GBS V2 exhibited the highest heterozygosity, minor allele frequency, and proportion of paralogs. More SNPs were uniquely identified by Stacks than TASSEL, though there was high overlap between methods. Researchers studying other conifer species should be prepared to analyze very large numbers of SNPs, and to consider the benefits and limitations of different pipelines.

2.1 Introduction

Single Nucleotide Polymorphisms (SNPs) have been widely used for plant genomic studies, including genome-wide association studies, marker-assisted breeding and genomic selection, because of their abundance in the genomes and amenability to high-throughput, cost effective genotyping technologies (Eckert et al. 2009, Hufford et al. 2012, Morris et al. 2013). Genotyping-by-Sequencing (GBS), which can generate tens of thousands of SNP markers without the need for a reference genome, has emerged as a cost-effective strategy for genome-wide SNP discovery and genotyping (Elshire et al. 2011, Andrews et al. 2016). By combining the power of multiplexed next-generation sequencing (NGS) with restriction enzyme based genome complexity reduction, GBS is able to genotype large populations of individuals for many thousands of SNPs in an increasingly rapid and inexpensive way (Poland et al. 2012, Poland and Rife 2012). Moreover, GBS has the potential to reach regions of the genome involved in transcription

regulation that are inaccessible to sequence capture approaches that target coding sequences (Mammadov et al. 2012).

However, genotyping and identifying SNPs is a challenge in most conifer species, due to their extremely large (19-32 Gb) and highly repetitive genomes. GBS has been tested for conifers on small numbers of individuals (<10) and has been found to produce tens of thousands of SNPs with high coverage (Chen et al. 2013, Pan et al. 2015). However, the use of GBS on conifers species is still largely limited by the difficulty of genome-wide SNP calling from the massively parallel short-read sequences (Glenn 2011, Goto et al. 2017). Even though GBS only sequences a fraction of the genome, because conifer genomes are so large and repetitive the datasets produced still present a computational challenge.

Studies now commonly use advanced analysis pipelines to filter, sort, and align the GBS raw data to get SNP data. There are two general types of pipelines for handling GBS data: reference-based and *de novo* approaches. Reference-based pipelines require an available reference genome, and call SNPs by mapping the raw GBS data to the reference genome to identify the position of sequences and compare the sequences from the same position to call SNPs (Nielsen et al. 2011). Several reference-based pipelines have been widely used, including: TASSEL-GBS (v1 and v2), Stacks, IGST, and Fast-GBS (Sonah et al. 2013, Catchen et al. 2013, Glaubitz et al. 2014, Torkamaneh et al. 2017). In the absence of a reference genome, *de novo* pipelines identify pairs of nearly identical reads (presumed to represent alternative alleles of a locus) to call SNPs. Two *de novo* pipelines are commonly used: the Universal Network Enabled Analysis Kit (UNEAK) (Lu et al. 2013), and Stacks (Catchen et al. 2013).

Previous studies have generally found that alignment to a reference genome from the same species increases the number of identifiable SNPs compared to the *de novo* pipelines (Torkamaneh et al. 2016). However, it is unknown which pipeline is best for SNP calling in species that lack a sequenced genome. This includes conifer species, most of which have as yet no available sequenced genome (Birol et al. 2013, Zimin et al. 2014, Stevens et al. 2016). Though aligning sequences to the reference genome of a closely related species could allow for more SNPs to be identified if sequences are fairly conserved, it could also result in many sequence fragments being rejected (and therefore SNPs in these fragments not being identified) if this is not the case.

No reference genome is available for ponderosa pine (*Pinus ponderosa*), but one does exist for loblolly pine (*Pinus taeda)* (Zimin et al. 2014, Neale et al. 2014). Of the conifers that have been sequenced to date, *P. taeda* is the most closely related to *P. ponderosa*; both are classified in the subgenus *Pinus* as opposed to *Strobus* (Gernandt et al. 2009, Willyard et al. 2009), which contains the other sequenced pine, *P. lambertiana* (Stevens et al. 2016). Furthermore, the *P. taeda* reference genome was successfully used to design probes for sequence capture in *P. contorta* (Suren et al. 2016, Yeaman et al. 2016). Recent studies show that within this subgenus, *P. taeda* and *P. ponderosa* diverged more recently from each other than either did from lodgepole pine (*P. contorta*) (Gernandt et al. 2005, Eckert and Hall 2006), suggesting that there is likely substantial sequence similarity between *P. taeda* and *P. ponderosa* as well. Previous studies have used *de novo* pipelines such as UNEAK to identify >10,000 SNP loci in conifers that lack a full genome sequence (Chen et al. 2013, Pan et al. 2015). However, these earlier studies

were based on a small number of samples, usually six individuals. Inclusion of more individuals will likely increase the number of SNPs identified – but by how much, and will the inclusion of more individual-level variation change the relative efficiency of different pipelines?

Despite the many advantages of GBS data, its reliability for SNP calling is compromised by the presence of paralogous genomic regions. Especially for the large genomes of conifers, involving both polyploidy and repetitive element activity (Li et al. 2015), it is challenging to separate multiple copies in a genome (e.g. paralogs) from variants at a single locus due to sequence similarity and the short sequences obtained. Moreover, it is largely unknown which pipeline does a better job at filtering out the paralogs.

In this study, we sequenced 94 individual *P. ponderosa* using GBS and compared four pipelines for SNP calling, including two reference based pipelines (TASSEL-GBS V2, Stacks), and two *de novo* pipelines (UNEAK, Stacks). We first tested the performance of various restriction enzymes for fragmentation of *P. ponderosa* genome, and then used the best for GBS library construction. Then we applied the TASSEL-GBS V2 (Glaubitz et al. 2014) and Stacks (Liu and Stützel 2004, Catchen et al. 2013) pipelines using the reference genome of *P. taeda*, as well as the Stacks and UNEAK (Lu et al. 2013) pipelines without a reference genome. Our aim was to determine which method produced the most SNPs, which produced the least amount of missing data for the SNPs identified, and how much overlap there is in the SNPs called between methods, as well as the proportion of paralogs among the SNPs called by different pipelines.

2.2 Materials and methods

2.2.1 Sample preparation

In the 1970s, the Forest Service's Pacific Southwest Regional Genetic Resources Program planted clones of 302 wild ponderosa pines in Chico, California. They came from diverse climate conditions in the central portion of California's Sierra Nevada mountains and are now reproductively mature, thus presenting an excellent resource for genetic studies. Although *P. ponderosa* contains multiple subdivisions, with the most important being between the Pacific and Rocky Mountain groups, based on their source locations the trees within the orchard likely do not cross any subdivision boundaries (Conkle and Critchfield 1988, Burns and Honkala 1990, Willyard et al. 2009, Potter et al. 2015).

For this study, we chose 94 individual *P. ponderosa* genotypes from the orchard collection. The source locations of these 94 genotypes are shown in Figure 2.1. The sample preparation includes three steps: dry needle preparation, DNA extraction, and quantification. Fresh needles were collected and dried with silica gel desiccant. Total genomic DNA was extracted from the dried needle tissue using DNeasy Plant Mini Kit (250) following the protocol from the manufacturer (Qiagen, Hilden, Germany) with two main modifications. First, to reduce protein contamination, for the step of grinding needles we added 1.5 ul of Proteinase K (20mg/ml) along with the Buffer AP and RNase A. The MiniG 1600 from SPEX SampePrep (Metuchen, NJ, USA) was used to grind needles with automated mechanical disruption through bead beating. Second, at the very last step, the amount of AE elution buffer was changed from 100 µl to 50 µl to get a

higher concentration of DNA (averagely 200 ng/µl). The DNA concentration was quantified using an Eppendorf BioSpectrometer (Eppendorf, AG, Germany).

2.2.2 Restriction enzyme selection

When working on a new species, it is beneficial to determine which enzyme produces the most fragments within the desired size range (100-400 bp). For optimization of the GBS protocol, 1000 ng samples of *P. ponderosa* genomic DNA were digested separately with *Ape*KI, *Pst*I, and *Eco*T22I, and with a combination of *Pst*I and *Eco*T22I (double digest) following the instructions of the enzyme manufacturer (New England Biolabs). These three restriction enzymes are methylation sensitive and have been previously used for construction of reduced complexity GBS libraries in conifers (Chen et al. 2013, Pan et al. 2015). Fragment size distributions of each test library were visualized using an Agilent BioAnalyzer 2100 (Agilent Technologies, Santa Clara, CA) with the High Sensitivity DNA Kit for quantification. For each test library, we have three samples with the same DNA concentration (50 ng/ul). We selected the enzyme based on the smoothness of the distribution and the size of the fragment sequences produced. Once this was done, all of the post-extraction steps were carried out at the UC Davis Genome Center.

2.2.3 Illumina libraries preparation and sequencing

For *Pinus contorta* and *Picea glauca* 47-plex GBS libraries yielded good results (Chen et al. 2013). Therefore, a 48-plex GBS library consisting of 47 DNA samples and a negative control (no DNA) was prepared in our study. The GBS protocol was slightly modified from the standard protocol (Elshire et al. 2011) and that of Chen et al. (2013). The library preparation and sequencing includes 6 steps: digestion, ligation, pooling samples, PCR, clean-up, and single-end read sequencing. DNA extracts (100 ng) were digested with the restriction enzyme *Ape*KI at 75 °C for 2 hours. Each of the 47 ponderosa pine DNA samples was tagged with a unique barcode. Sequences for the *Ape*KI barcode adapters and the common adapters, and the temperature cycles, were as described in Chen et al. (2013). After the digestion, the samples were cooled to 4 °C, and then adapters were ligated onto restriction fragments. This was done using T4 DNA Ligase (Life Technologies, Burlington, ON, Canada) at 16 °C for 1 hour, after which samples were "heat killed" at 65 °C for 20 minutes. The pool was quantified via qPCR using the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA) for Illumina sequencing platforms, with 0.9X bead cleanup to remove small fragments (<250 bp). Additional DNA purification using the Zymo DNA Clean & Concentrator kit (Zymo Research, Irvine, CA) was performed to further increase the purity of the extracted DNA. The fragments from all 47 samples were then sequenced (single-end read 90 bp) on one lane of an Illumina HiSeq 4000 (Illumina, San Diego, CA). The same procedure was repeated for the other 47 samples (single-end read 100 bp). We then assessed the sequence quality of the raw reads using FastQC analysis (Simon 2010).

2.2.4 SNP calling

We used the reference genome of loblolly pine v2.0 (https://treegenesdb.org/FTP/Genomes/Pita/) for the reference-based pipelines. TASSEL-

GBS V2 is implemented in TASSEL V5.0, a program originally developed for maize to facilitate genotype-phenotype comparisons (https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Tassel5GBSv2Pipeline). This pipeline requires a reference genome to call SNPs. The steps involved are illustrated on the left side of Figure 2.2. The raw sequence data in FASTQ file are first trimmed to same length (64 bp) and then identical reads are assembled into tags (unique DNA sequences). These distinctive tags are saved into FASTQ file. Then alignment program bowtie2 (Langmead and Salzberg 2012) is used to align these tags with the reference genome of loblolly pine. Based on the position of the tags against the reference genome, SNPs are produced by identifying tags aligned in the same position that have a 1 bp mismatch. Finally, the SNP information within each tag for each sample are output as a VCF file. Each step is performed internally with TASSEL-GBS V2 plugins except for alignment, which is carried out externally with bowtie2. We used the default parameter settings for our analysis except that the minimum quality score was set to 20 to make the base call accuracy more than 99%.

Stacks is a software package developed for restriction site-associated DNA sequencing that identifies SNPs and calculates population statistics from any restriction enzyme-based, reduced-representation sequence data with short-read sequences (http://catchenlab.life.illinois.edu/stacks/). It was developed with population genomics in mind, and so aims to assemble loci in large numbers of individuals and read haplotypes from them. Stacks allows for SNP calling with or without a reference genome; we chose to do both. The details steps of Stacks reference pipeline are represented on the right side of Figure 2.2. There are two main differences with the TASSEL-GBS V2 pipeline. First, Stacks reference pipeline aligns the reads directly against the reference genome, while TASSEL-GBS V2 pipeline assembles the same reads into tags and then performs the alignment. Second, the BWA alignment program (Li and Durbin 2009) is used instead of bowtie2. Each step in the Stacks reference pipeline is performed internally in Stacks algorithms except alignment with BWA and the SAMtools (Li 2011) step used to get read position.

The steps involved in the Stacks *de novo* pipeline are shown on the right side of Figure 2.3. First, reads are demultiplexed, cleaned and trimmed to 64 bp, and identical reads are assembled as "stacks". The stacks in each sample are merged as catalogs, which then are grouped together across samples. Third, SNPs are identified by matching reads to the catalogs and assigning SNPs to each sample when there is a 1 bp mismatch. SNP information is saved in a VCF file. Optional additional steps include the creation of genetic maps and calculation of population statistics. Every step in Stacks *de novo* pipeline uses the Stacks internal algorithms. For both Stacks reference and *de novo* pipeline, we used the default parameter settings except that the quality score limit was set to 20 instead of 15, for greater accuracy and to be consistent with TASSEL-GBS V2.

The UNEAK (Universal Network Enabled Analysis Kit) pipeline can be implemented in TASSEL V3.0 (https://tassel.bitbucket.io/TasselArchived.html), but it is not available in V5.0. UNEAK is a *de novo* pipeline that can call SNPs without a reference genome. The steps in the UNEAK pipeline are on the left side of Figure 2.2. The general design of UNEAK is as follows: 1) raw Illumina DNA sequence data were first trimmed to 64-bp; 2) identical 64-bp reads for each individual are collapsed into tags; 3) pair-wise alignment identifies tag pairs having a single base pair mismatch. These

single base pair mismatches are candidate SNPs, which are then assigned to each sample and saved as VCF file. As in the Stacks *de novo* pipeline, every step in UNEAK pipeline uses the internal algorithms. We again used the default parameter settings except that the base call accuracy is changed from 0.03 to 0.01, which is equivalent to the first two methods.

Due to the large genome size of pines, the raw data for each of the two sets of 47 samples was over 19 GB after compression. Large computing resources are needed to run these pipelines. We ran most of the steps on the Multi-Environment Research Computer for Exploration and Discovery (MERCED) cluster, a shared resource for UC Merced researchers, which has 128 GB RAM in each compute node. The exception was the step cstacks (Figure 2.3, merge stacks as catalogs) in the Stacks *de novo* pipeline, which requires a very large memory and RAM. For this, we used the XSEDE supercomputing resource (Towns et al. 2014), which has 3000 GB RAM in each computer node.

2.2.5 SNP quality and comparison

To evaluate the quality of the SNPs in each VCF output file, six parameters were chosen: good reads, the missing genotype rate, minor allele frequency (MAF), heterozygosity, read depth, and the proportion of paralogs. We used PLINK 1.9 (Purcell and Chang 2015), a widely used open-source C/C++ toolset in population genetics, to calculate the missing genotype rate, MAF, and heterozygosity. Whole-genome duplications have occurred in conifers (Li et al. 2015, Prunier et al. 2016), resulting in multiple paralogs. Such paralogs could yield false SNPs if incorrectly identified as a single locus based on short GBS sequence reads. To address this issue and distinguish real allelic variation from paralogs, we tested for deviations in ratio of read depth for each allele within heterozygotes in the GBS data (McKinney et al. 2017). The deviation of this ratio from its expected value (1:1) is expressed as a Z-score with a binomial distribution ($P = 0.5$). Based on these Z-scores, we declare likely paralog variants using a conservative threshold of $|Z| > 5$.

Besides the quality of the SNPs, we were also interested in how many SNPs were identified by more than one pipeline. In our study, the comparison of SNP overlap was done using VCFtools (Danecek et al. 2011).

2.3 Results

2.3.1 Restriction enzyme selection

Figure 2.4 shows the amplified fragment size distributions of libraries from ponderosa pine DNA digested with different restriction enzymes. *Ape*KI yielded a high smooth curve of fragment sizes between 150 and 500, which indicates good performance for GBS. *Pst*I performed similarly, though the curve was more jagged. EcoT22I produced a lower, more jagged curve, while EcoT22I + P*st*I had the worst performance of all. We therefore selected *Ape*KI. This enzyme does not cut CpG methylated sequences (Castel et al. 2011), and therefore tends to avoid stably silenced portions of the genome, hopefully increasing the proportion of SNPs from more actively transcribed regions.

2.3.2 Sequence quality of raw reads

Quality control for the raw reads involves the analysis of sequence quality, GC content, sequence length distribution, the presence of adaptors, overrepresented sequences, sequence duplication levels in order to detect sequencing errors, PCR artifacts, or contamination. Reducing the error rate of base calls and improving the accuracy of the per-base quality score are integral to having reliable GBS raw data (Nielsen et al. 2011). The sequence quality of the raw reads was high, with the per base sequence quality score over 32 and the most frequently observed mean quality score per sequence over 40. This indicates that the sequencing error is less than 0.1%.

The per base GC content module measures the GC content across the whole length of each sequence in a file and compares it to a modeled normal distribution of GC content. For the raw data of our study, the per base GC content is a roughly normal distribution with both the shape and peak corresponding to the distribution of GC content of the underlying genome, which indicates a normal random library without any bias.

According to the sequence length distribution plot, the length for all the sequences is, as expected, 90 bp for one set of samples and 100 bp for the other. The duplicate sequences analysis issues an error since non-unique sequences make up more than 50% of the total, which is in line with the high proportions of repetitive sequences in conifers (Morse et al. 2009, Kovach et al. 2010). This is a feature that could be problematic for SNP calling; however, each pipeline has its own method for cleaning the data that can be more or less effective at removing repetitive sequences. No sequence represented more than 0.1% of the total, indicating that the library was not contaminated.

2.3.3 Comparison of four SNP-calling pipelines

The four SNP-calling pipelines differed in many respects (Table 2.1). Of the two *de novo* pipelines, Stacks identified fewer SNPs than UNEAK (62,882 vs. 196,698) and took much longer to run than any of the other three pipelines. Of the two reference-based analyses, Stacks identified 25% more SNPs than TASSEL-GBS V2 and took about 57% longer to run. The two reference-based pipelines identified over an order of magnitude more SNPs than the two *de novo* pipelines. For the Stacks pipeline, the reference-based version identified over forty times as many SNPs as the *de novo* one with a shorter run time.

The SNP quality data includes good reads, missing data, average MAF, and average observed and expected heterozygosity, average read depth per individual, and the proportion of paralogs. There were 7.8 billion total reads for the 94 samples. All the five pipelines used the same quality score (20) and same length (64 bp) to clean and trim the raw data. However, the number of reads considered "good" differed between pipelines, with TASSEL-GBS V2 keeping only 76.9% of reads, while the others kept at least 93.6% (Table 2.1). This resulted in TASSEL-GBS V2 having a much lower missing genotype rate (47.4% vs >72%). The TASSEL-GBS V2 pipeline produced the largest average read depth per individual (22.5 vs. < 5). The relatively low read depth of Stacks reference-based pipeline (5.8) and Stacks *de novo* pipeline (4.6) is consistent with their high percentages of missing genotype calls.

The UNEAK pipeline produced a much smaller average MAF than the other pipelines (0.093 vs. > 0.21). This is likely due to UNEAK employing a network filter to discard repeats and paralogs. Accordingly, UNEAK produced a small proportion of

paralogs (1.1%). The proportion of paralogs of TASSEL-GBS V2 pipeline is much higher than the other pipelines (18.5% vs. < 1.5%). The higher numbers of SNPs identified by TASSEL-GBS V2 pipeline is partly due to paralogs.

Interestingly, reference-based Stacks identified a very low average observed heterozygosity despite having SNPs with a relatively high minor allele frequency. Stacks *de novo* and TASSEL-GBS V2 had similar minor allele frequencies and expected heterozygosity, but the observed heterozygosity was higher for TASSEL-GBS V2. For all pipelines, the average observed heterozygosity is lower than expected heterozygosity, which suggests that at least some loci are out of Hardy Weinberg equilibrium. This may be due to selection or genetic drift operating across the Sierra Nevada mountains, as the sampled individuals are widely distributed and do not represent a single random-mating population.

There are 1,888,913 overlapping SNPs identified by the two reference-based pipelines (Figure 2.5). Of the SNPs identified by TASSEL-GBS V2 11.4% were unique, while of those identified by the Stacks reference pipeline 30.2% were unique. The vcf-compare function compares SNPs based on their position relative to the loblolly pine genome. Because the positions of SNPs were identified based on the reference genome, we were only able to compare the SNPs found using the two reference-based pipelines. Efforts to map SNPs identified by the *de-novo* approaches to the genome were stymied by the fact that the loblolly genome has not been fully assembled into chromosomes, and we were not able to develop a work-around for this that would enable software like VCFtools to be used.

2.4 Discussion

The repetitive DNA content in conifers affects the efficiency of SNP calling (Pan et al. 2015) and requires strategies for reducing the complexity and repetitive DNA content of GBS libraries. Selection of a restriction enzyme (RE) is one of the critical steps in GBS (Elshire et al. 2011, Peterson et al. 2012). In our study, the commonly-used restriction enzyme *Ape*KI performed well for ponderosa pine, with *Pst*I offering a decent second choice. As shown in Figure 2.3, there were no discrete peaks suggesting repetitive DNA fragments present in *Ape*KI library, while the other three REs had a few discrete peaks. GBS libraries derived from *Ape*KI also had a higher proportion of fragments within the sequencing size range (<500bp). Similarly, for lodgepole pine (*Pinus contorta*) and white spruce (*Picea glauca*), Chen et al. (2013) found that the size distribution curve was smoothest for *Ape*KI compared to *Pst*I and *Eco*T22I. *Ape*KI was also used for other conifers such as interior spruce, a hybrid complex of white spruce (*Picea glauca*) and Engelmann spruce (*Picea engelmannii*) (Gamal El-Dien et al. 2015). Thus, *Ape*KI seems to be a good choice for conifers in general.

As Table 2.2 indicates, no one pipeline was superior in all categories that might be of concern for a researcher. Both the reference based and *de novo* SNP calling approaches work for ponderosa pine, but the reference based pipelines using the loblolly pine genome identified more SNPs with a reasonable level of coding complexity and computing resources (Table 2.1). This suggests that for other non-model species without available genome sequences, SNP calling using a reference genome from closely related species can be an effective option. This difference can be explained by their alignment

methods. All these pipelines assemble identical reads as tags/stacks before the alignment. The reference-based pipelines then align the tags/stacks with the reference genome to find their position, and then compare the tags/stacks in the same positions to identify SNPs with 1 bp mismatch (Figure 2.1). Thus, the reference genome helps to ensure that tags from the same position are compared to identify SNPs. The *de novo* pipelines directly compare the tags/stacks with each other to identify SNPs with 1 bp mismatch (Figure 2.3). In this situation, some of the reads from the same general position may not be identified as pairs because not enough of their short sequences overlap, and therefore some of the SNPs are missed. Torkamaneh et al. (2016) conducted a comparison between different SNP calling pipelines on soybean (*Glycine max*) and found that four reference-based pipelines (TASSEL-GBS V1, IGST, TASSEL-GBSV2, Fast-GBS) identified more SNPs than either of two *de novo* pipelines (Stacks, UNEAK). However, the differences between the methods were much smaller than the differences found in our study.

Even within the two *de novo* pipelines, the number of SNPs identified were very different. Torkamaneh et al. (2016) also found that the UNEAK pipeline identified more SNPs than the Stacks *de novo* pipeline. One possible explanation for this difference is the different way of assembling the identical reads as tags/stacks. For Stacks, the default setting for the maximum number of stacks at a single *de novo* locus in the program ustacks is three. If there are over three stacks in the same locus, it will be blacklisted, meaning that locus will not be available for insertion into, or matching against, the catalogue. This is done as a means of rejecting repetitive sequences. However, the UNEAK merges the identical reads as tags without this limit. As a result, UNEAK pipeline can potentially identify most of SNPs because fewer stacks are rejected, but could also have more errors involving not properly separating paralogs. However, as discussed below, this did not appear to be the case; the percentage of paralogs was similar. Given this, and the more efficient identification of SNPs, we would recommend UNEAK over Stacks for *de novo* SNP identification.

The different number of SNPs identified by the two reference-based pipelines is likely caused by a difference in how they assemble tags/stacks. TASSEL-GBS V2 assembles the identical reads as tags first, and then align the tags to the reference genome. Stacks aligns the trimmed reads directly to the reference genome, which may lead to more alignments and a greater number of SNPs identified. The Stacks reference pipeline ran slower but identified more SNPs than TASSEL-GBS V2. All the steps in TASSEL-GBS V2 could deal with all the 94 samples together and assign the SNPs data into each sample in the final VCF file. However, some steps in Stacks (e.g. ustacks, SAMtools) need to have separate codes for each sample instead of the one code for 47 samples together as a group, which takes more effort. TASSEL-GBS V2 rejected a higher proportion of reads initially (lower % considered "good") but produced a much lower percentage of missing data by either locus or individual than the other methods, which would mean less imputation will be needed at later steps in an association or genetic structure analysis. However, despite this thinning of reads, TASSEL-GBS V2 appears to be more likely to incorrectly identify SNPs from paralogs than the other three methods. Thus, for reference-based assembly, we would again recommend Stacks based on lower paralog percentages and higher SNP number, with the caveat that it is somewhat less user-friendly.

There are 1,888,931 SNPs identified by both of the reference-based methods. These SNPs comprised most (88.6%) of those identified by TASSEL-GBS V2. This pipeline exhibited the highest heterozygosity, MAF and proportion of paralogs, so some of the loci identified that did not overlap (11.4%) likely had unusually high heterozygosity. Stacks, which produced the highest number of loci, did so in part by identifying 816,107 SNPs that were not identified by TASSEL-GBS V2.

Finally, while earlier studies making use of <10 individual conifers identified <20,000 SNPs (Chen et al. 2013, Pan et al. 2015), this study identified between 62,882 and 2,705,038 SNPs from 94 individuals. This indicates the high degree of genetic variation that is present in ponderosa pine (Potter et al. 2015) and within other widespread conifer species (Potter et al. 2012). While these individuals came from multiple populations within the Sierra Nevada, this represents only a tiny fraction of the total range of this species, which extends from northern Mexico to southern Canada and from the Pacific to the Rocky Mountains. Future studies, especially those considering range-wide variation, should be prepared to analyze very high numbers of SNPs.

2.5 Acknowledgements

We thank the Forest Service's Pacific Southwest Regional Genetic Resources Program for allowing us to sample needles from their seed orchard. The sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. For SNP identification, we made use of the MERCED computer cluster at UC Merced (supported by NSF Award ACI-1429783) and the Extreme Science and Engineering Discovery Environment (XSEDE; supported by NSF Award ACI-1548562). We also thank Stephen Hart, Jeffrey Lauder and Melaine Aubry-Kientz for their comments on this manuscript.

2.6 References

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature reviews. Genetics 17:81–92.

Birol, I., A. Raymond, S. D. Jackman, S. Pleasance, R. Coope, G. A. Taylor, M. M. S. Yuen, C. I. Keeling, D. Brand, B. P. Vandervalk, H. Kirk, P. Pandoh, R. A. Moore, Y. Zhao, A. J. Mungall, B. Jaquish, A. Yanchuk, C. Ritland, B. Boyle, J. Bousquet, K. Ritland, J. MacKay, J. Bohlmann, and S. J. M. Jones. 2013. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497.

Burns, R. M., and B. H. Honkala. 1990. Silvics of North America. USDA Forest Service, Washington DC.

Castel, A. L., M. Nakamori, C. A. Thornton, and C. E. Pearson. 2011. Identification of restriction endonucleases sensitive to 5-cytosine methylation at non-CpG sites, including expanded (CAG)n/(CTG)n repeats. Epigenetics 6:416–420.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124–3140.

Chen, C., S. E. Mitchell, R. J. Elshire, E. S. Buckler, and Y. A. El-Kassaby. 2013. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genetics & Genomes 9:1537–1544.

Conkle, M. T., and W. B. Critchfield. 1988. Genetic variation and hybridization of ponderosa pine. Ponderosa pine: the species and its management 27:43.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Eckert, A. J., A. D. Bower, J. L. Wegrzyn, B. Pande, K. D. Jermstad, K. V. Krutovsky, J. B. St. Clair, and D. B. Neale. 2009. Association Genetics of Coastal Douglas Fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-Hardiness Related Traits. Genetics 182:1289–1302.

Eckert, A. J., and B. D. Hall. 2006. Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): Phylogenetic tests of fossil-based hypotheses. Molecular Phylogenetics and Evolution 40:166–182.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:1–10.

Gamal El-Dien, O., B. Ratcliffe, J. Klápště, C. Chen, I. Porth, and Y. A. El-Kassaby. 2015. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics 16:370.

Gernandt, D. S., G. Geada Lopez, S. O. Garcia, and A. Liston. 2005. Phylogeny and classification of *Pinus*. Taxon 54:29–42.

Gernandt, D. S., S. Hernández-León, E. Salgado-Hernández, and J. A. Pérez de La Rosa. 2009. Phylogenetic relationships of Pinus subsection Ponderosae inferred from rapidly evolving cpDNA regions. Systematic Botany 34:481–491.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, Q. Sun, and E. S. Buckler. 2014. TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11:759–769.

Goto, S., H. Kajiya-Kanegae, W. Ishizuka, K. Kitamura, S. Ueno, Y. Hisamoto, H. Kudoh, M. Yasugi, A. J. Nagano, and H. Iwata. 2017. Genetic mapping of local adaptation along the altitudinal gradient in Abies sachalinensis. Tree Genetics & Genomes 13:104.

Hufford, M. B., X. Xu, J. van Heerwaarden, T. Pyhäjärvi, J.-M. Chia, R. A. Cartwright, R. J. Elshire, J. C. Glaubitz, K. E. Guill, S. M. Kaeppler, J. Lai, P. L. Morrell, L. M. Shannon, C. Song, N. M. Springer, R. A. Swanson-Wagner, P. Tiffin, J. Wang, G. Zhang, J. Doebley, M. D. McMullen, D. Ware, E. S. Buckler, S. Yang, and J. Ross-Ibarra. 2012. Comparative population genomics of maize domestication and improvement. Nature Genetics 44:808–811.

Kovach, A., J. L. Wegrzyn, G. Parra, C. Holt, G. E. Bruening, C. A. Loopstra, J. Hartigan, M. Yandell, C. H. Langley, I. Korf, and D. B. Neale. 2010. The Pinus taeda

genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11:420.
Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357–359.
Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England) 27:2987–2993.
Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
Li, Z., A. E. Baniaga, E. B. Sessa, M. Scascitelli, S. W. Graham, L. H. Rieseberg, and M. S. Barker. 2015. Early genome duplications in conifers and other seed plants. Science Advances 1:e1501084.
Lu, F., A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney, M. D. Casler, E. S. Buckler, and D. E. Costich. 2013. Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. PLOS Genetics 9:e1003215.
Mammadov, J., R. Aggarwal, R. Buyyarapu, and S. Kumpatla. 2012. SNP Markers and Their Impact on Plant Breeding. Research article. https://www.hindawi.com/journals/ijpg/2012/728398/.
McKinney, G. J., R. K. Waples, L. W. Seeb, and J. E. Seeb. 2017. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Molecular Ecology Resources 17:656–669.
Morris, G. P., P. Ramu, S. P. Deshpande, C. T. Hash, T. Shah, H. D. Upadhyaya, O. Riera-Lizarazu, P. J. Brown, C. B. Acharya, S. E. Mitchell, J. Harriman, J. C. Glaubitz, E. S. Buckler, and S. Kresovich. 2013. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences 110:453–458.
Morse, A. M., D. G. Peterson, M. N. Islam-Faridi, K. E. Smith, Z. Magbanua, S. A. Garcia, T. L. Kubisiak, H. V. Amerson, J. E. Carlson, C. D. Nelson, and J. M. Davis. 2009. Evolution of Genome Size and Complexity in Pinus. PLOS ONE 4:e4332.
Neale, D. B., J. L. Wegrzyn, K. A. Stevens, A. V. Zimin, D. Puiu, M. W. Crepeau, C. Cardeno, M. Koriabine, A. E. Holtz-Morris, J. D. Liechty, P. J. Martínez-García, H. A. Vasquez-Gross, B. Y. Lin, J. J. Zieve, W. M. Dougherty, S. Fuentes-Soriano, L.-S. Wu, D. Gilbert, G. Marçais, M. Roberts, C. Holt, M. Yandell, J. M. Davis, K. E. Smith, J. F. Dean, W. W. Lorenz, R. W. Whetten, R. Sederoff, N. Wheeler, P. E. McGuire, D. Main, C. A. Loopstra, K. Mockaitis, P. J. deJong, J. A. Yorke, S. L. Salzberg, and C. H. Langley. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology 15:R59.
Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12:443–451.

Pan, J., B. Wang, Z.-Y. Pei, W. Zhao, J. Gao, J.-F. Mao, and X.-R. Wang. 2015. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Molecular Ecology Resources 15:711–722.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLOS ONE 7:e37135.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J.-L. Jannink. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLOS ONE 7:e32253.

Poland, J. A., and T. W. Rife. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome 5:92–102.

Potter, K. M., V. D. Hipkins, M. F. Mahalovich, and R. E. Means. 2015. Nuclear genetic variation across the range of ponderosa pine (Pinus ponderosa): Phylogeographic, taxonomic and conservation implications. Tree Genetics & Genomes 11.

Potter, K. M., R. M. Jetton, W. S. Dvorak, V. D. Hipkins, R. Rhea, and W. A. Whittier. 2012. Widespread inbreeding and unexpected geographic patterns of genetic variation in eastern hemlock (Tsuga canadensis), an imperiled North American conifer. Conservation Genetics 13:475–498.

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.

Purcell, S., and C. Chang. 2015. PLINK 1.9. URL https://www. cog-genomics. org/plink2.

Simon, A. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Sonah, H., M. Bastien, E. Iquira, A. Tardivel, G. Légaré, B. Boyle, É. Normandeau, J. Laroche, S. Larose, M. Jean, and F. Belzile. 2013. An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLoS ONE 8:1–9.

Stevens, K. A., J. L. Wegrzyn, A. Zimin, D. Puiu, M. Crepeau, C. Cardeno, R. Paul, D. Gonzalez-Ibeas, M. Koriabine, A. E. Holtz-Morris, P. J. Martínez-García, U. U. Sezen, G. Marçais, K. Jermstad, P. E. McGuire, C. A. Loopstra, J. M. Davis, A. Eckert, P. de Jong, J. A. Yorke, S. L. Salzberg, D. B. Neale, and C. H. Langley. 2016. Sequence of the Sugar Pine Megagenome. Genetics 204:1613–1626.

Suren, H., K. A. Hodgins, S. Yeaman, K. A. Nurkowski, P. Smets, L. H. Rieseberg, S. N. Aitken, and J. A. Holliday. 2016. Exome capture from the spruce and pine giga-genomes. Molecular Ecology Resources 16:1136–1146.

Torkamaneh, D., J. Laroche, M. Bastien, A. Abed, and F. Belzile. 2017. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. BMC Bioinformatics 18:5.

Torkamaneh, D., J. Laroche, and F. Belzile. 2016. Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. PLoS ONE 11.

Towns, J., T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkens-Diehr.

2014. XSEDE: Accelerating scientific discovery. Computing in Science and Engineering 16:62–74.
Willyard, A., R. Cronn, and A. Liston. 2009. Reticulate evolution and incomplete lineage sorting among the ponderosa pines. Molecular Phylogenetics and Evolution 52:498–511.
Yeaman, S., K. A. Hodgins, K. E. Lotterhos, H. Suren, S. Nadeau, J. C. Degner, K. A. Nurkowski, P. Smets, T. Wang, L. K. Gray, K. J. Liepe, A. Hamann, J. A. Holliday, M. C. Whitlock, L. H. Rieseberg, and S. N. Aitken. 2016. Convergent local adaptation to climate in distantly related conifers. Science 353:1431–1433.
Zimin, A., K. A. Stevens, M. W. Crepeau, A. Holtz-Morris, M. Koriabine, G. Marçais, D. Puiu, M. Roberts, J. L. Wegrzyn, P. J. de Jong, D. B. Neale, S. L. Salzberg, J. A. Yorke, and C. H. Langley. 2014. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics 196:875–890.

Table 2.1 Comparison of different SNP-calling approaches.

Approach	*de novo*		reference-based	
Pipeline	Stacks	UNEAK	TASSEL-GBS V2	Stacks
Run time (hours: min)	53:26	2:17	21:8	33:45
Number of good reads (billion)	7.5	7.3	6.0	7.5
Percent of good reads (%)	96.2	93.6	76.9	96.2
Total SNPs	62,882	196,698	2,131,362	2,705,038
Missing data (%)	72.3	73.9	47.4	76.0
Average MAF	0.275	0.093	0.273	0.217
Observed heterozygosity	0.258	0.044	0.306	0.066
Expected heterozygosity	0.334	0.147	0.348	0.288
Average read depth per individual (Standard Deviation)	13.2 (2.2)	4.6 (0.6)	22.5 (5.5)	5.8 (1.0)
Paralogs (%)	1.5	1.1	18.5	1.0

Table 2.2 SNP-calling approaches ranked

Approach	*de novo*		reference-based	
Pipeline	Stacks	UNEAK	TASSEL-GBS V2	Stacks
Run time	Highest	Lowest	Medium	Medium
Ease of use	Poor	Best	Medium	Poor
# of SNPs identified	Lowest	Low	High	Highest
Missing data	High	High	Lowest	High
% paralogs	Low (good)	Low (good)	Highest (poor)	Lowest (best)

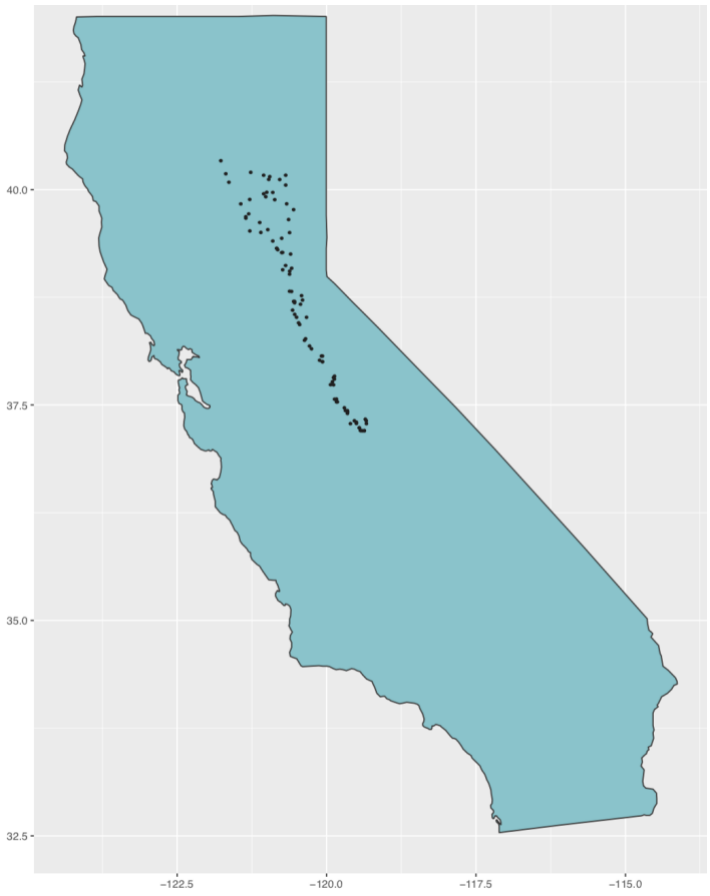

Figure 2.1 Geographic distribution of the 94 samples. The black dots represent original genotype source locations.

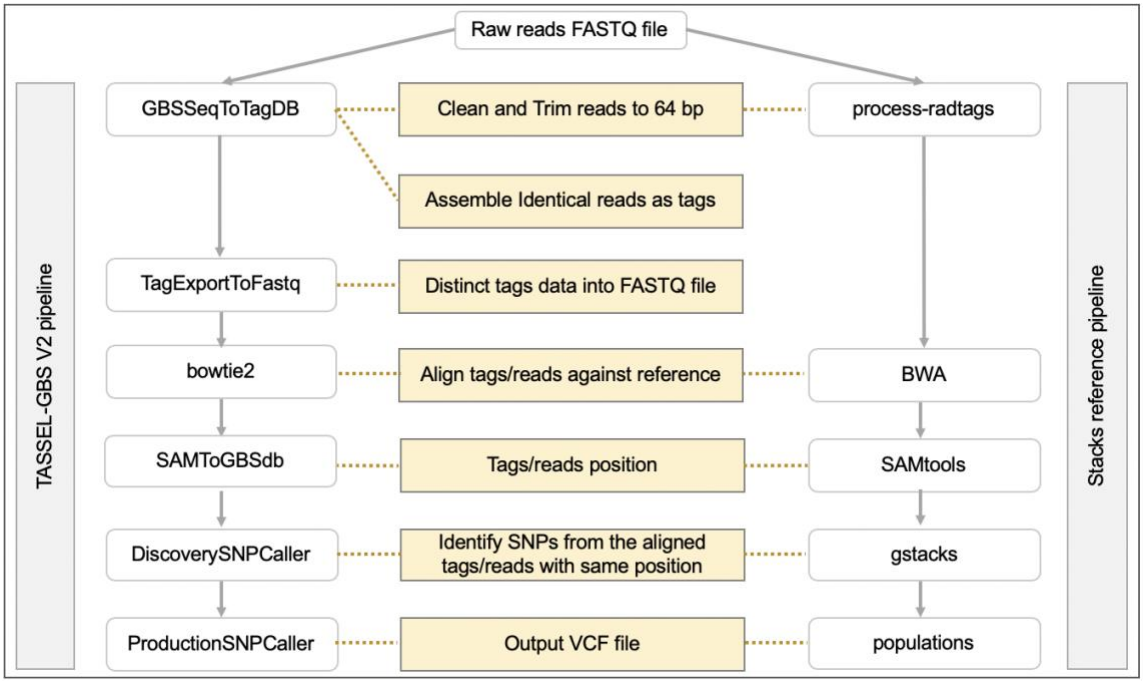

Figure 2.2 Comparison of the two reference-based pipelines. The horizontal boxes on the left side represent the programs in GBS V2. The horizontal boxes on the right side represent the programs in the Stacks reference pipeline. The yellow boxes in the middle represent potential program functions, while the yellow dotted lines specify the main function for each program in the two pipelines.

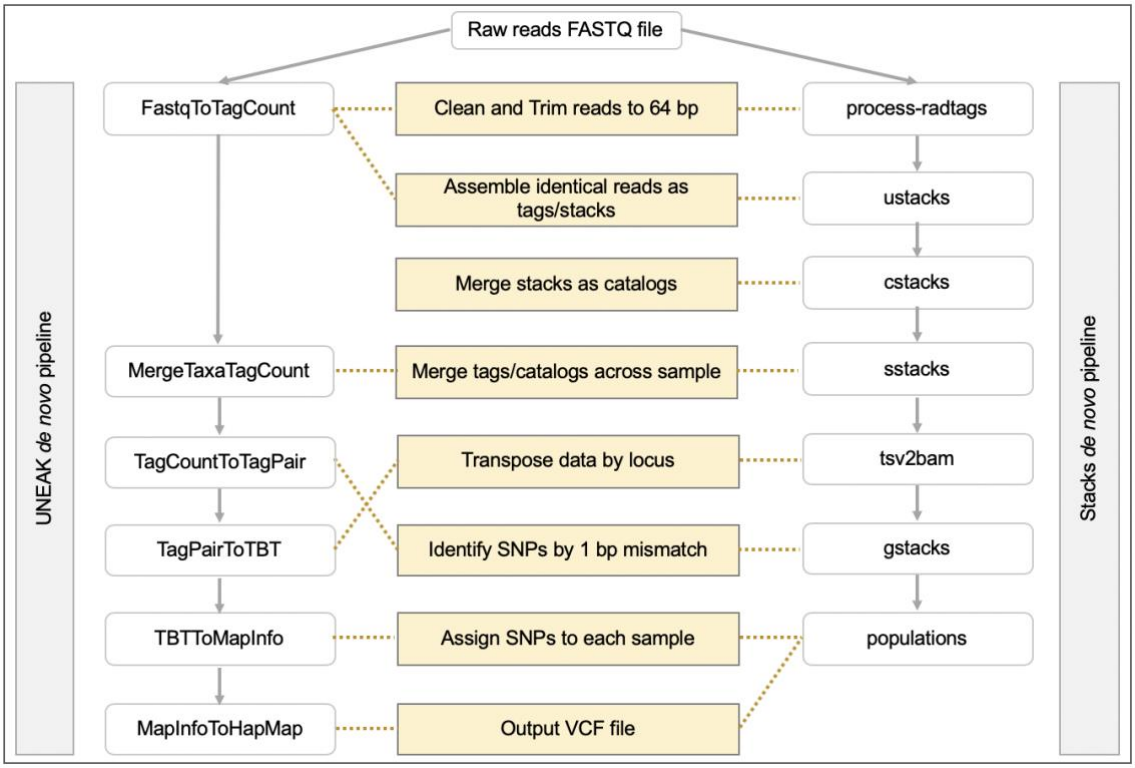

Figure 2.3 Comparison between two *de novo* pipelines. The horizontal boxes on the left side represent the programs in UNEAK *de novo*. The horizontal boxes on the right side represent the programs in Stacks *de novo*. The yellow boxes in the middle represent the functions of the program, while the yellow dotted lines specify the main function for each program in the two pipelines.

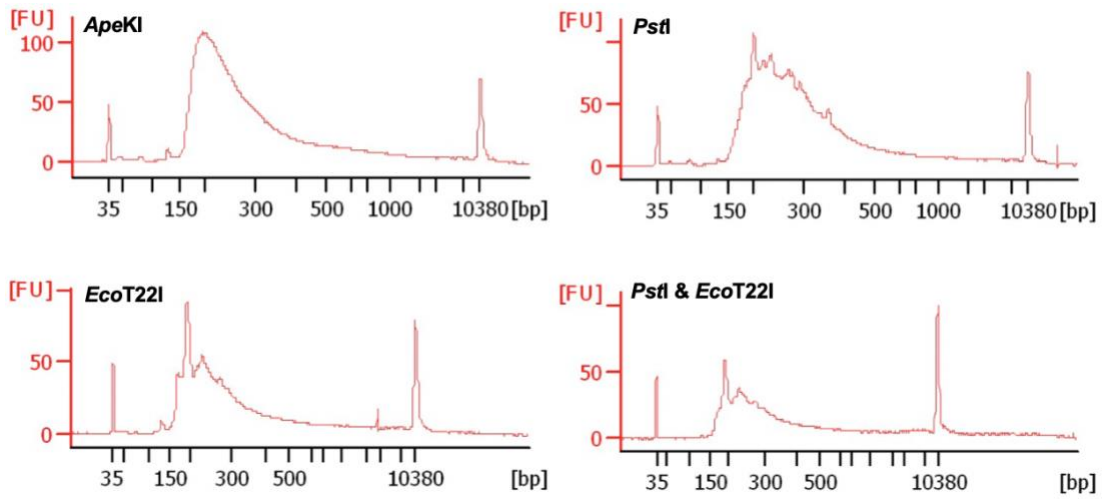

Figure 2.4 Fragment size distribution of GBS libraries with different restriction enzymes. The y-axis shows fluorescence units, indicating amount of DNA. Numbers below hatch marks on the x-axis indicate fragment size (bp).

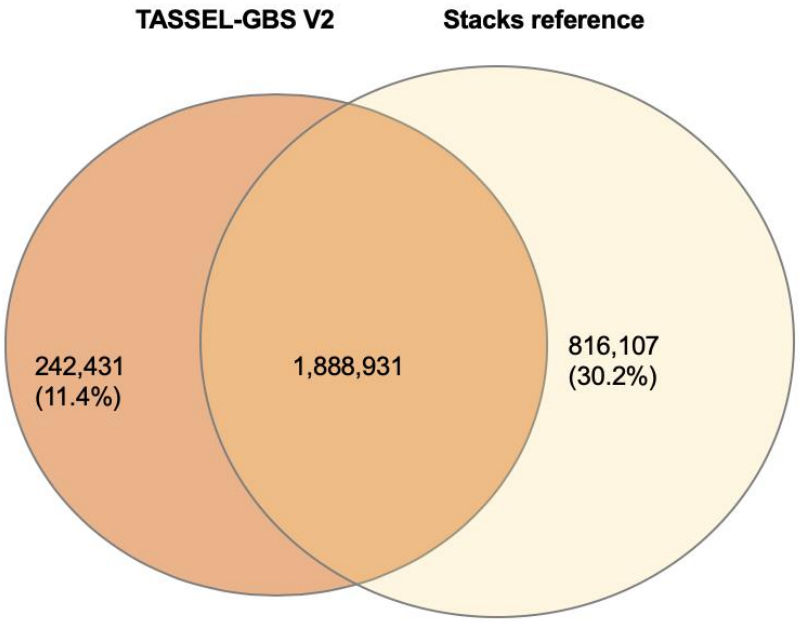

Figure 2.5 Venn diagram comparing SNPs overlap between the two reference-based pipelines. The circle on the left side represents the SNPs produced by TASSEL-GBS V2 pipeline. The circle on the right side represents the SNPs produced by Stacks reference-based pipeline.

Chapter 3:
Identifying environmentally associated genetic variation in ponderosa pine

3.0 Abstract

A fundamental goal of evolutionary biology is to understand adaptive genetic variation and the concomitant evolutionary potential of a species. Genotype-to-environment (G2E) association analysis has enormous potential to discover genes responsible for local adaptation. In non-model species, genotyping by sequencing (GBS), a reduced representation sequencing approach, generates large numbers of single nucleotide polymorphisms (SNPs) in an efficient and inexpensive way. Sequences from 223 *Pinus ponderosa* (ponderosa pine) individuals were aligned to the reference genome of *Pinus taeda* (loblolly pine), a closely related species, to identify SNPs. We ran a G2E analysis with these SNPs and five chosen climatic variables using LFMM2, which controls for the effects of demographic processes and population structure on the distribution of genetic variation. We found 213 SNPs strongly associated with mean maximum summer temperature (TMAX), 335 with mean minimum winter temperature (TMIN), 1798 with April 1st snow pack (PCK4), and 120 SNPs with mean climatic water deficit (CWD). No SNPs were associated with mean monthly winter precipitation (PPTW). Different protein functions have been annotated underlying the genetic associations, including ubiquitination, abscisic acid (ABA) signaling pathway, cell division or growth of roots or shoots, cell wall organization, seed dormancy. Potentially, future studies can develop molecular tools based on the associated genetic markers, which are necessary to understand trees' adaptative responses to environmental variation, to assist breeders and gene resource managers in developing and managing adapted populations.

3.1 Introduction

Understanding adaptive genetic variation and the concomitant evolutionary potential of a species is a central aim in conservation and evolutionary biology (Hoffmann and Sgrò 2011, Savolainen et al. 2013, Harrisson et al. 2014). Microevolution – that is, adaptation within species and populations via changes in allele frequencies or genotypic recombination – can be important for local species persistence under environmental change (Bell and Gonzalez 2009). Intraspecific genetic variation represents the potential for further adaptive change in response to new selective challenges such as global warming (Rice and Emery 2003). A long history of studies in forestry have clearly demonstrated the existence of local adaptation in tree populations (Langlet 1971, Ying and Liang 1994, Kitzmiller 2005, Wright 2007). However, locally adapted tree populations with long life cycles may become maladapted if climate-induced environmental shifts outpace range shifts, plastic responses, or evolutionary adaptation (Aitken et al. 2008, Anderson et al. 2012, Alberto et al. 2013). Studies examining species distribution models indicate that not accounting for regional adaptive genetic variation could result in downward-biased predictions of species distribution areas under climate

change (Garzón et al. 2011). Understanding the distribution of genetic variation related to environmental responses may help us better predict and manage forests in a changing climate (Neale and Kremer 2011, Oney et al. 2013).

Landscape genomics, which investigates the statistical association between genetic variation at individual loci and environmental gradients, offers enormous potential to discover genes responsible for local adaptation (Eckert et al. 2010, 2015, Sork et al. 2013, Lu et al. 2019). This approach is sometimes known as genotype-to-environment (G2E) association analysis (Chapter 1). However, the investigation of local adaptation at a genetic level is still limited in non-model species, especially for forest trees (Neale and Kremer 2011, Bragg et al. 2015). Most of the studies that have taken this approach in trees focused on a modest number of candidate genes (Holliday et al. 2010, Hamilton et al. 2013, Dillon et al. 2014). Though targeted sequencing for candidate genes is efficient, they may miss other important genes or regulatory regions with previously unsuspected roles in local adaptation. The sequencing of a large number of genome-wide genetic markers can help to uncover such missing genes. To achieve this, several approaches based on next generation sequencing (NGS) have been proposed in recent years (Davey et al. 2011, Poland and Rife 2012). Genotyping-by-Sequencing (GBS), which can generate tens of thousands of SNP markers (Single Nucleotide Polymorphisms) without the need for a reference genome, has emerged as a cost-effective strategy (Elshire et al. 2011, Andrews et al. 2016). By combining the power of multiplexed NGS with restriction enzyme based genome complexity reduction, GBS is able to genotype large populations of individuals for many thousands of SNPs in an increasingly rapid and inexpensive way (Poland et al. 2012, Poland and Rife 2012). GBS has been found produce tens of thousands of SNPs with high coverage in conifers from 10 or fewer individuals (Chen et al. 2013, Pan et al. 2015).

When conducting G2E studies, it is important to control for the effects of demographic processes and population structure on the distribution of genetic variation (Wang et al. 2017). Approaches to deal with this include BAYENV (Günther and Coop 2013), BAYPASS (Gautier 2015), BAYESCENV (Villemereuil and Gaggiotti 2015), and latent factor mixed model (LFMM) (Frichot et al. 2013, Frichot and François 2015). These models can effectively account for population structure and can accommodate large SNP data sets. However, these methods rely on Markov chain Monte Carlo algorithms or Bayesian bootstrap methods to perform parameter inference and statistical testing, which can be computationally intensive and slow. One method, LFMM2, was developed for G2E association and has been shown to outperform other similar approaches with several orders-of-magnitude faster computing (Caye et al. 2019).

Despite the high economic and ecological importance of ponderosa pine (*Pinus ponderosa*) in the western United States, no previous study has investigated the genetics of drought tolerance in this species, or conducted a G2E analysis. Some studies have investigated *P. ponderosa* evolutionary history and phylogeography using mitochondrial DNA markers; these reflect long-term biogeographical process (e.g. glacial isolation) contributed to the modern distribution of the evolutionary lineage, but have little adaptive significance in themselves (Johansen and Latta 2003, Potter et al. 2013). Other studies have emphasized the importance of intraspecific variation of *P. ponderosa* in adaptation to climate change, but mainly focus on the phenotypic variation within and among

populations without identifying the underlying genetic variation (Kolb et al. 2016b, Maguire et al. 2018).

For California, regional climate change models predict an increase in temperature and summer drought periods (IPCC 2014). California's historic 2012–2016 drought may represent an increasingly common condition in which high temperatures coincide with a series of dry years (Griffin and Anchukaitis 2014, Berg and Hall 2015). One of the most prominent effects of this hot drought has been an increase in conifer mortality in the Sierra Nevada, which may negatively impact the sustainability of conifer forests (Fettig et al. 2019). A deep understanding of the genetic basis of adaptation in ponderosa pine is critical for successful reforestation, for conservation and restoration programs, and for managing or predicting climate-induced species range changes.

In this study, we use SNPs derived from GBS of widely distributed individuals of ponderosa pine from the Sierra Nevada to run a G2E analysis in combination with gene annotation. We dissect the genetic variants associated with different climate variables to identify loci potentially related to local adaptation to environmental conditions in ponderosa pine.

3.2 Materials and Methods

3.2.1 Sampling and sequencing

In the 1970s, the Forest Service's Pacific Southwest Regional Genetic Resources Program planted clones of 302 wild ponderosa pines in Chico, California. They came from diverse climate conditions in the central portion of California's Sierra Nevada and are now reproductively mature, thus presenting an excellent resource for genetic studies. For this study, we chose 223 individual *P. ponderosa* genotypes from the orchard collection. The source locations of these 223 genotypes are shown in Figure 3.1. These locations likely fall within just one of the several genetic subdivisions previously identified in ponderosa pine (Conkle and Critchfield 1988, Willyard et al. 2009b, Potter et al. 2015).

The sample preparation includes dry needle preparation, DNA extraction, and quantification (Chapter 2). After the DNA extraction, four sets of a 48-plex GBS library consisting of 47 DNA samples and a negative control (no DNA) and one set of a 36-plex GBS library consisting of 35 DNA samples and a negative control were prepared by the UC Davis Genome Center, with the methods described in Chapter 2. The pool was quantified via qPCR using the KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA) for Illumina sequencing platforms, with 0.9X bead cleanup to remove small fragments (<250 bp). Additional DNA purification using the Zymo DNA Clean & Concentrator kit (Zymo Research, Irvine, CA) was performed to further increase the purity of the extracted DNA. The fragments from 47 samples at a time were then sequenced (single-end read 90 bp or 100 bp) on one lane of an Illumina HiSeq 4000 (Illumina, San Diego, CA).

3.2.2 SNP calling

No reference genome is available for ponderosa pine (*Pinus ponderosa*), but one does exist for loblolly pine (*Pinus taeda)* (Zimin et al. 2014, Neale et al. 2014). Of the conifers that have been sequenced to date, *P. taeda* is the most closely related to *P.*

ponderosa (Gernandt et al. 2009, Willyard et al. 2009a). Furthermore, the *P. taeda* reference genome was successfully used to design probes for sequence capture in *P. contorta* (Suren et al. 2016, Yeaman et al. 2016). Based on the analyses in Chapter 2, we selected the Stack v.2.2 pipeline (Rochette and Catchen 2017) with this reference genome (https://treegenesdb.org/FTP/Genomes/Pita/) for SNP calling. Each step in the Stacks reference pipeline is performed internally in Stacks algorithms except alignment with BWA v.0.7.17 (Li and Durbin 2009) and the Samtools v.1.9 (Li 2011) step used to get read position. Default settings were used in Stacks, BWA and Samtools.

3.2.3 Climate data

We obtained 30-year (1951–1980) averages of climate data from the 270 m resolution California Basin Characterization Model (BCM) (Flint et al. 2013). We selected five variables for G2E analysis that had low-to-moderate correlations with each other (Chapter 1). These five variables include: mean climatic water deficit (CWD, a measure of evaporative demand exceeding soil moisture); mean minimum winter temperature (TMIN; December–February); mean maximum summer temperature (TMAX; June–August); mean monthly winter precipitation (PPTW;December–February); and April 1st snow pack (PCK4). Temperature and precipitation are among the major ecological variables that determine plants' natural distribution and drive their adaptation (Berry and Bjorkman 1980).

3.2.4 Environmental associations

In this study, we choose LFMM 2 (Caye et al. 2019) to run the genotype to environment association analysis. LFMM2 regression models combine fixed and latent effects with the following equation:

$$\mathbf{Y}=\mathbf{X}\mathbf{B}^{T}+\mathbf{W}+\mathbf{E}.$$

where **Y** is a matrix of genetic information measured from *p* genetic markers for *n* individuals, and the variables of interest (environment variables), **X**, measured for *d* environmental variables and n individuals. The fixed effect sizes are recorded in the B matrix, which has dimension *p* * *d*. The **E** matrix represents residual errors with the same dimensions as the response matrix. The matrix **W** is a latent matrix of rank K, defined by K latent factors where K can be determined by model choice procedures. The K latent factors represent unobserved confounders - usually geographical structure in the genotypes of the samples - which are modeled through a *n**K matrix, **U**. **V** is a p × K matrix of loadings. The matrix U is obtained from a singular value decomposition (SVD) of the matrix.

$$\mathbf{W}=\mathbf{U}\mathbf{V}^{T}$$

To determine K, we used principal component analysis (PCA) and admixture analysis as implemented in the LEA v.2.6.0 R package (Frichot et al. 2013, Frichot and François 2015). First, we ran the LEA function pca to compute the scores of a PCA and select the number of significant components by computing Tracy-Widom tests with the LEA function tracy.widom (Patterson et al. 2006). Second, we ran the LEA function snmf for each K value between 1 and 5 with 10 repetitions. By comparing the cross-entropies as described in the snmf manual (Frichot & Francois, 2014), the most likely K

value was identified by minimizing the cross-validation error evaluated in the 10-fold cross-validation procedure. After we run LFMM2, we then chose significant associations based on a false rate of 5% ($q \leqslant 0.05$) using the R package QVALUE (Storey and Tibshirani 2003).

LFMM approaches are robust to high amounts of missing data, such as GBS sequencing tends to produce, when sample sizes are >100 (Xuereb et al. 2017). For this analysis, we chose to focus on a subset of raw environmental variables (eg. maximum summer temperature) rather than environmental PCA axes, as a number of previous studies have done (eg. Eckert et al. 2010, 2015). We did this because PCA associations can be difficult to interpret if, for example, both PCA axes include temperature and moisture variables, and because we were able to identify five environmental variables that had low to moderate correlation with each other and which might be associated with different adaptive responses.

3.2.5 Gene annotation

After we got the significantly associated SNPs, we ran SnpEff (Cingolani et al. 2012) for SNP annotation. We built the data base with the annotated genome and the reference genome of loblolly pine v.2.01 in TreeGenes (http://treegenesdb.org/FTP/Genomes/Pita/v2.01/). Then we aligned the gene sequence against the nonredundant protein sequences database using UniProt to identify the gene and protein with the implemented Blastx (V2.9.0, $e < 1e{-}10$). The Gene Ontology Annotation Database ("UniProt" 2015, Bateman et al. 2017) was used to further identify the potential functions of the genes.

3.3 Results

3.3.1 Genetic diversity and population structure

After SNP calling and filtering, a total of 4,155,896 SNPs remained for PCA and LFMM2 analysis. According to the PCA results with all SNPs, two principle components explained the genetic variation between our samples (Table 3.1). The best K value based on LEA snmf is one (Figure 3.2). We also plotted the admixture of each individual tree using the snmf results and found no signal of two populations (Figures 3.3 & 3.4). Thus, we assumed that these 223 individuals belong to one interbreeding population and ran LFMM 2.0 using K = 1.

3.3.2 Environmental associations at individual loci

After the running of LFMM2 ($q \leq 0.05$), we found many significant associations between SNPs and the environmental variables. There are 213 SNPs strongly associated with TMAX, 335 with TMIN, 1798 with PCK4, and 120 SNPs with CWD. However, no SNP was found to be significantly with PPTW. There were 62 SNPs associated with both PCK4 and TMIN, 45 associated with both CWD and TMIN, 7 SNPs associated with both TMAX and PCK4, and 1 SNP associated with both PCK4 and CWD (Figure 3.5).

3.3.3 Gene annotation

The location of each SNP is listed in the output file of SnpEff. Accordingly, there are mainly six location categories, including intragenic variants, intergenic variants,

upstream SNPs, downstream SNPs, synonymous, and missense variants in the gene coding sequence. In SnpEff, "intragenic" refers to SNPs in introns rather than exons of genes, while "missense" refers to any non-synonymous mutation in the transcribed region. As shown in Table 3.2, most of the SNPs are between genes (in the intergenic regions) and likely have no direct effect on gene expression.

For the gene annotation, we focused on the other five types of SNP variant. We found several protein types that are likely relevant to drought tolerance and other environmental responses (Table 3.3). Some of the SNPs associated with TMAX (maximum summer temperature), TMIN (minimum winter temperature), CWD (climatic water deficit), and PCK4 (April snowpack) are in or near genes in the jasmonic acid synthesis or response pathways and the protein ubiquitination pathway, both of which are associated with responses to biotic or abiotic stress. Climatic water deficit and snowpack were also associated with SNPs in or near genes involved in seed dormancy and the abscisic acid (ABA) signaling pathway, both of which have been previously linked to drought responses in trees (Moran et al. 2017). Genes involved in reproduction, including pollen and ovule formation, were associated with TMAX, TMIN, and PCK4. CWD and PCK4 were associated with genes involved in cell wall organization. Both TMAX and PCK4 were associated with genes involved in xylem and phloem formation, and growth regulation and stress responses, while TMIN and PCK4 were associated with genes involved in stomatal regulation and pathogen responses. Further biotic and abiotic stress response genes were associated with PCK4, as were genes involved in nutrient transport, photosynthesis, respiration, sugar synthesis, and light responses.

For many of the other loci associated with environmental gradients, gene ontology results were too vague to draw many conclusions about their function or why the association might exist. However, some of these genes have been previously associated with stress, including Ras-related protein RABC1 to drought responses (Khassanova et al. 2019), and pentatricopeptide repeat-containing protein to cold stress (Xing et al. 2018). Two of the SNPs associated with minimum temperature are found in the intragenic region of CGS1 and RE2, genes known to be upregulated during cold stress (Dinari et al. 2013) and heat stress (Traylor-Knowles et al. 2017), respectively. Most of the others are involved in gene expression (RNA or DNA binding, transcription factors, helicase activity, ribosome components, methylation) or ATP binding.

3.4 Discussion

Based on the gene ontology and protein function annotation of climate-associated SNPs, we found many linked to genes implicated in abiotic or biotic stress responses. We also identified several SNPs in or near genes with previously unsuspected roles in local adaptation. Our findings suggest the efficiency of G2E analysis with GBS to uncover the adaptive genetic variation in ponderosa pine as well as the important genes and proteins involved, thus provide new insights on the adaptive potential of ponderosa pine.

Over half (1729) of the SNPs identified as being associated with climate were only associated with PCK4. This may reflect the importance of snow affecting the distribution of ponderosa pine in the Sierra Nevada mountains. In this Mediterranean climate region, most of the annual precipitation occurs during the winter, and the spring and early summer discharge of its major rivers is supplied mostly by melting of winter

snow accumulation at high elevations (Serreze et al. 1999). However, a heavy snowpack may also delay the start of the growing season for juvenile trees. Consistent with this, at least one of the associated SNPs was linked to light response. The number of SNPs associated with more than one climatic variable was surprisingly low, but this may simply indicate that we were successful in selecting climatic variables that are not strongly correlated with one another and which require different genetic adaptations in *P. ponderosa*. The highest degree of overlap was between PCK4 and TMIN (62 SNPs) and between CWD and TMIN (45 SNPs). The former might be related to adaptation to cold versus heat and/or winter precipitation, while the latter might be related to how quickly the site warms up, drying out the soil.

Most of the environmentally associated SNPs were identified as being intergenic and so their function, if any, is unclear. Of the remaining SNPs, most of them are in introns, which are not transcribed. There are also many synonymous mutations, which do not result in an amino acid change and are assumed to be neutral with respect to fitness. Either intragenic or synonymous variants might also be in linkage disequilibrium with a causal variant outside of the sequenced area. There were also quite a few upstream and downstream SNPs that could affect gene expression. Finally, non-synonymous variants may directly affect phenotype as they alter which amino acid is coded for; 53 of the climate-associated SNPs fell into this category.

The prevalence of genetic associations related to ubiquitination and abscisic acid-signaling pathways is consistent with prior studies of drought response in conifers (Moran et al. 2017). Increasing abscisic acid (ABA) concentrations are used as a signal to keep stomata closed during dry conditions, reducing water loss (Brodribb et al. 2014). In addition, ABA signaling can also affect shoot growth and water uptake (Buckley 2005, Hamanishi and Campbell 2011). Ubiquitin has been found to be involved in drought responses in model species by playing a role in ABA-mediated dehydration stress responses (Ryu et al. 2010, Kim et al. 2012), or through the downregulation of plasma membrane aquaporin levels (Lee et al. 2009). The study of the role of ubiquitin in conifer drought response is still somewhat limited.

Quite a few genes involved with cell division or growth of roots or shoots were found to be associated with April snowpack and maximum temperature, which are consistent with the conifer growth patterns. The main burst of growth in Sierra Nevada conifers begins with the melting of the snowpack in the spring while the soil is moist from winter precipitation (Royce and Barbour 2001). Thus, April snowpack can act as a selective force for the genes involved in cell division or growth in conifers. Moreover, root and shoot morphology may also strongly affect how ponderosa pine access and use water during the summer hot and dry period (Kolb et al. 2016a). Summer temperatures are relevant for drought stress, as higher temperatures mean faster evaporation of water from the soil. In seedlings of Sierra Nevada conifers, maximum summer temperature has a strong effect on growth and survival (Moran et al. 2019). Many of the same genes associated with these climatic factors have also been linked to various biotic and abiotic stress responses. These stresses can interact with, for example, drought stress being associated with greater risk of bark beetle attack in pines (Kane and Kolb 2010, Fettig et al. 2019).

SNPs in or around genes involved in vascular tissue formation were found to be associated with maximum temperature or snowpack, which may be related to managing water transport in hotter and drier versus wetter and colder environments. We hypothesize that the genes involved in cell wall organization and associated with CWD or snowpack might also be involved in xylem formation, as the degree of cell expansion during xylem formation can have a strong effect on the drought resistance of that xylem (Anfodillo et al. 2012, Bryukhanova and Fonti 2013). Various genes linked to reproduction were associated with April snowpack and maximum and minimum temperatures. Pines release their pollen in early spring, and so early spring temperatures could affect reproductive success, while summer stress could impede cone growth. signaling can affect shoot growth and water uptake.

To conclude, by investigating adaptive genetic variation in ponderosa pine with G2E association analysis, our study has found numerous genomic variants distributed across the genome with gene function associated with response to climate. With the identified associations, it is possible to develop molecular tools based on the associated genetic markers to assist breeders and gene resource managers in developing and managing adapted populations, which have been lacking to date. These tools may thus contribute to a way to shorten the long periods of time that tree breeders need to assess adaptation; this effort is especially important given the current rapidly changing climate. In addition, our results should open up new opportunities for functional studies to determine the molecular roles of the genes underlying these associated genetic makers in influencing trees adaptation.

3.5 References

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. Evolutionary Applications 1:95–111.

Alberto, F. J., S. N. Aitken, R. Alía, S. C. González-Martínez, H. Hänninen, A. Kremer, F. Lefèvre, T. Lenormand, S. Yeaman, R. Whetten, and O. Savolainen. 2013. Potential for evolutionary responses to climate change – evidence from tree populations. Global Change Biology 19:1645–1661.

Anderson, J. T., A. M. Panetta, and T. Mitchell-Olds. 2012. Evolutionary and Ecological Responses to Anthropogenic Climate Change: Update on Anthropogenic Climate Change. Plant Physiology 160:1728–1740.

Andrews, K. R., J. M. Good, M. R. Miller, G. Luikart, and P. A. Hohenlohe. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature reviews. Genetics 17:81–92.

Anfodillo, T., A. Deslauriers, R. Menardi, L. Tedoldi, G. Petit, and S. Rossi. 2012. Widening of xylem conduits in a conifer tree depends on the longer time of cell expansion downwards along the stem. Journal of Experimental Botany 63:837–845.

Bateman, A., M. J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-A-Jee, A. Cowley, A. D. Silva, M. D. Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y.

Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, A. Renaux, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.-C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.-L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L.-S. Yeh, and J. Zhang. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45:D158–D169.

Bell, G., and A. Gonzalez. 2009. Evolutionary rescue can prevent extinction following environmental change. Ecology Letters 12:942–948.

Berg, N., and A. Hall. 2015. Increased Interannual Precipitation Extremes over California under Climate Change. Journal of Climate 28:6324–6334.

Berry, J., and O. Bjorkman. 1980. Photosynthetic Response and Adaptation to Temperature in Higher Plants. Annual Review of Plant Physiology 31:491–543.

Bragg, J. G., M. A. Supple, R. L. Andrew, and J. O. Borevitz. 2015. Genomic variation across landscapes: insights and applications. New Phytologist 207:953–967.

Brodribb, T. J., S. A. M. McAdam, G. J. Jordan, and S. C. V. Martins. 2014. Conifer species adapt to low-rainfall climates by following one of two divergent pathways. Proceedings of the National Academy of Sciences 111:14489–14493.

Bryukhanova, M., and P. Fonti. 2013. Xylem plasticity allows rapid hydraulic adjustment to annual climatic variability. Trees 27:485–496.

Buckley, T. N. 2005. The control of stomata by water balance. New Phytologist 168:275–292.

Caye, K., B. Jumentier, J. Lepeule, and O. François. 2019. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Molecular Biology and Evolution 36:852–860.

Chen, C., S. E. Mitchell, R. J. Elshire, E. S. Buckler, and Y. A. El-Kassaby. 2013. Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genetics & Genomes 9:1537–1544.

Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6:80–92.

Collevatti, R. G., E. Novaes, O. B. Silva-Junior, L. D. Vieira, M. S. Lima-Ribeiro, and D. Grattapaglia. 2019. A genome-wide scan shows evidence for local adaptation in a widespread keystone Neotropical forest tree. Heredity 123:117–137.

Conkle, M. T., and W. B. Critchfield. 1988. Genetic variation and hybridization of ponderosa pine. In: Ponderosa Pine: the species and its management, Washington State University Cooperative Extension, 1988: p. 27-43.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12:499–510.

Dillon, S., R. McEvoy, D. S. Baldwin, G. N. Rees, Y. Parsons, and S. Southerton. 2014. Characterisation of Adaptive Genetic Diversity in Environmentally Contrasted Populations of Eucalyptus camaldulensis Dehnh. (River Red Gum). PLOS ONE 9:e103515.

Dinari, A., A. Niazi, A. R. Afsharifar, and A. Ramezani. 2013. Identification of Upregulated Genes under Cold Stress in Cold-Tolerant Chickpea Using the cDNA-AFLP Approach. PLoS ONE 8.

Eckert, A. J., J. van Heerwaarden, J. L. Wegrzyn, C. D. Nelson, J. Ross-Ibarra, S. C. Gonzalez-Martinez, and D. B. Neale. 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., P. E. Maloney, D. R. Vogler, C. E. Jensen, A. D. Mix, and D. B. Neale. 2015. Local adaptation at fine spatial scales: an example from sugar pine (Pinus lambertiana, Pinaceae). Tree Genetics & Genomes 11.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:1–10.

Fettig, C. J., L. A. Mortenson, B. M. Bulaon, and P. B. Foulk. 2019. Tree mortality following drought in the central and southern Sierra Nevada, California, U.S. Forest Ecology and Management 432:164–178.

Flint, L. E., A. L. Flint, J. H. Thorne, and R. Boynton. 2013. Fine-scale hydrologic modeling for regional landscape applications: the California Basin Characterization Model development and performance. Ecological Processes 2:1–21.

Frichot, E., and O. François. 2015. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6:925–929.

Frichot, E., S. D. Schoville, G. Bouchard, and O. François. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular Biology and Evolution 30:1687–1699.

Garzón, M. B., R. Alía, T. M. Robson, and M. A. Zavala. 2011. Intra-specific variability and plasticity influence potential tree species distributions under climate change. Global Ecology and Biogeography 20:766–778.

Gautier, M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics 201:1555–1579.

Gernandt, D. S., S. Hernández-León, E. Salgado-Hernández, and J. A. Pérez de La Rosa. 2009. Phylogenetic relationships of Pinus subsection Ponderosae inferred from rapidly evolving cpDNA regions. Systematic Botany 34:481–491.

Griffin, D., and K. J. Anchukaitis. 2014. How unusual is the 2012–2014 California drought? Geophysical Research Letters 41:2014GL062433.

Günther, T., and G. Coop. 2013. Robust Identification of Local Adaptation from Allele Frequencies. Genetics 195:205–220.
Hamanishi, E. T., and M. M. Campbell. 2011. Genome-wide responses to drought in forest trees. Forestry: An International Journal of Forest Research 84:273–283.
Hamilton, J. A., C. Lexer, and S. N. Aitken. 2013. Differential introgression reveals candidate genes for selection across a spruce (Picea sitchensis × P. glauca) hybrid zone. New Phytologist 197:927–938.
Harrisson, K. A., A. Pavlova, M. Telonis-Scott, and P. Sunnucks. 2014. Using genomics to characterize evolutionary potential for conservation of wild populations. Evolutionary Applications 7:1008–1025.
Hoffmann, A. A., and C. M. Sgrò. 2011. Climate change and evolutionary adaptation. Nature 470:479–485.
Holliday, J. A., K. Ritland, and S. N. Aitken. 2010. Widespread, ecologically relevant genetic markers developed from association mapping of climate-related traits in Sitka spruce (Picea sitchensis). New Phytologist 188:501–514.
IPCC. 2014. Climate Change 2014: Impacts,Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Page 1132. Cambridge University Press, Cambridge, UK and New York, USA.
Johansen, A. D., and R. G. Latta. 2003. Mitochondrial haplotype distribution, seed dispersal and patterns of postglacial expansion of ponderosa pine. Molecular Ecology 12:293–298.
Kane, J. M., and T. E. Kolb. 2010. Importance of resin ducts in reducing ponderosa pine mortality from bark beetle attack. Oecologia 164:601–609.
Khassanova, G., A. Kurishbayev, S. Jatayev, A. Zhubatkanov, A. Zhumalin, A. Turbekova, B. Amantaev, S. Lopato, C. Schramm, C. Jenkins, K. Soole, P. Langridge, and Y. Shavrukov. 2019. Intracellular Vesicle Trafficking Genes, RabC-GTP, Are Highly Expressed Under Salinity and Rapid Dehydration but Down-Regulated by Drought in Leaves of Chickpea (Cicer arietinum L.). Frontiers in Genetics 10.
Kim, S. J., M. Y. Ryu, and W. T. Kim. 2012. Suppression of Arabidopsis RING-DUF1117 E3 ubiquitin ligases, AtRDUF1 and AtRDUF2, reduces tolerance to ABA-mediated drought stress. Biochemical and Biophysical Research Communications 420:141–147.
Kitzmiller, J. H. 2005. Provenance Trials of Ponderosa Pine in Northern California. Forest Science 51:595–607.
Kolb, T. E., K. C. Grady, M. P. McEttrick, and A. Herrero. 2016a. Local-Scale Drought Adaptation of Ponderosa Pine Seedlings at Habitat Ecotones. Forest Science 62:641–651.
Kolb, T. E., K. C. Grady, M. P. McEttrick, and A. Herrero. 2016b. Local-Scale Drought Adaptation of Ponderosa Pine Seedlings at Habitat Ecotones. Forest Science 62:641–651.
Lamara, M., G. J. Parent, I. Giguère, J. Beaulieu, J. Bousquet, and J. J. MacKay. 2018. Association genetics of acetophenone defence against spruce budworm in mature white spruce. BMC Plant Biology 18:231.

Langlet, O. 1971. Two Hundred Years Genecology. Taxon 20:653–721.
Lee, H. K., S. K. Cho, O. Son, Z. Xu, I. Hwang, and W. T. Kim. 2009. Drought Stress-Induced Rma1H1, a RING Membrane-Anchor E3 Ubiquitin Ligase Homolog, Regulates Aquaporin Levels via Ubiquitination in Transgenic Arabidopsis Plants. The Plant Cell 21:622–641.
Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England) 27:2987–2993.
Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.
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. Ecology and Evolution 9:6798–6809.
Maguire, K. C., D. J. Shinneman, K. M. Potter, and V. D. Hipkins. 2018. Intraspecific Niche Models for Ponderosa Pine (Pinus ponderosa) Suggest Potential Variability in Population-Level Response to Climate Change. Systematic Biology 67:965–978.
Moran, E. V., A. J. Das, J. Keeley, and N. L. Stephenson. 2019. Negative impacts of summer heat on Sierra Nevada tree seedlings. Ecosphere 10:e02776.
Moran, E. V., J. Lauder, C. Musser, A. Stathos, and M. J. Shu. 2017. The genetics of drought tolerance in conifers. New Phytologist 216:1034–1048.
Neale, D. B., and A. Kremer. 2011. Forest tree genomics: growing resources and applications. Nature Reviews Genetics 12:111–122.
Neale, D. B., J. L. Wegrzyn, K. A. Stevens, A. V. Zimin, D. Puiu, M. W. Crepeau, C. Cardeno, M. Koriabine, A. E. Holtz-Morris, J. D. Liechty, P. J. Martínez-García, H. A. Vasquez-Gross, B. Y. Lin, J. J. Zieve, W. M. Dougherty, S. Fuentes-Soriano, L.-S. Wu, D. Gilbert, G. Marçais, M. Roberts, C. Holt, M. Yandell, J. M. Davis, K. E. Smith, J. F. Dean, W. W. Lorenz, R. W. Whetten, R. Sederoff, N. Wheeler, P. E. McGuire, D. Main, C. A. Loopstra, K. Mockaitis, P. J. deJong, J. A. Yorke, S. L. Salzberg, and C. H. Langley. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biology 15:R59.
Oney, B., B. Reineking, G. O'Neill, and J. Kreyling. 2013. Intraspecific variation buffers projected climate change impacts on Pinus contorta. Ecology and Evolution 3:437–449.
Pan, J., B. Wang, Z.-Y. Pei, W. Zhao, J. Gao, J.-F. Mao, and X.-R. Wang. 2015. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Molecular Ecology Resources 15:711–722.
Patterson, N., A. L. Price, and D. Reich. 2006. Population Structure and Eigenanalysis. PLoS Genetics 2:e190.
Poland, J. A., P. J. Brown, M. E. Sorrells, and J.-L. Jannink. 2012. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLOS ONE 7:e32253.

Poland, J. A., and T. W. Rife. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome 5:92–102.
Potter, K. M., V. D. Hipkins, M. F. Mahalovich, and R. E. Means. 2013. Mitochondrial DNA haplotype distribution patterns in Pinus ponderosa (Pinaceae): range-wide evolutionary history and implications for conservation. American Journal of Botany 100:1562–1579.
Potter, K. M., V. D. Hipkins, M. F. Mahalovich, and R. E. Means. 2015. Nuclear genetic variation across the range of ponderosa pine (Pinus ponderosa): Phylogeographic, taxonomic and conservation implications. Tree Genetics & Genomes 11:38.
Rice, K. J., and N. C. Emery. 2003. Managing microevolution: restoration in the face of global change. Frontiers in Ecology and the Environment 1:469–478.
Rochette, N. C., and J. M. Catchen. 2017. Deriving genotypes from RAD-seq short-read data using Stacks. Nature Protocols 12:2640–2659.
Royce, E. B., and M. G. Barbour. 2001. Mediterranean climate effects. II. Conifer growth phenology across a Sierra Nevada ecotone. American Journal of Botany 88:919–932.
Ryu, M. Y., S. K. Cho, and W. T. Kim. 2010. The Arabidopsis C3H2C3-Type RING E3 Ubiquitin Ligase AtAIRP1 Is a Positive Regulator of an Abscisic Acid-Dependent Response to Drought Stress. Plant Physiology 154:1983–1997.
Savolainen, O., M. Lascoux, and J. Merilä. 2013. Ecological genomics of local adaptation. Nature Reviews Genetics 14:807–820.
Serreze, M. C., M. P. Clark, R. L. Armstrong, D. A. McGinnis, and R. S. Pulwarty. 1999. Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data. Water Resources Research 35:2145–2160.
Sork, V. L., S. N. Aitken, R. J. Dyer, A. J. Eckert, P. Legendre, and D. B. Neale. 2013. Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. Tree Genetics & Genomes 9:901–911.
Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100:9440–9445.
Suren, H., K. A. Hodgins, S. Yeaman, K. A. Nurkowski, P. Smets, L. H. Rieseberg, S. N. Aitken, and J. A. Holliday. 2016. Exome capture from the spruce and pine giga-genomes. Molecular Ecology Resources 16:1136–1146.
Traylor-Knowles, N., N. H. Rose, E. A. Sheets, and S. R. Palumbi. 2017. Early Transcriptional Responses during Heat Stress in the Coral Acropora hyacinthus. The Biological Bulletin 232:91–100.
UniProt: a hub for protein information. 2015. . Nucleic Acids Research 43:D204–D212.
Villemereuil, P. de, and O. E. Gaggiotti. 2015. A new FST-based method to uncover local adaptation using environmental variables. Methods in Ecology and Evolution 6:1248–1258.
Wang, J., Q. Zhao, T. Hastie, and A. B. Owen. 2017. CONFOUNDER ADJUSTMENT IN MULTIPLE HYPOTHESIS TESTING. Annals of statistics 45:1863–1894.
Willyard, A., R. Cronn, and A. Liston. 2009a. Reticulate evolution and incomplete lineage sorting among the ponderosa pines. Molecular Phylogenetics and Evolution 52:498–511.

Willyard, A., R. Cronn, and A. Liston. 2009b. Reticulate evolution and incomplete lineage sorting among the ponderosa pines. Molecular Phylogenetics and Evolution 52:498–511.
Wright, J. W. 2007. Local adaptation to serpentine soils in Pinus ponderosa. Plant and Soil 293:209–217.
Xing, H., X. Fu, C. Yang, X. Tang, L. Guo, C. Li, C. Xu, and K. Luo. 2018. Genome-wide investigation of pentatricopeptide repeat gene family in poplar and their expression analysis in response to biotic and abiotic stresses. Scientific Reports 8:1–9.
Xuereb, A., A. Stahlke, M. Bermingham, M. Brown, E. Nonaka, O. Razgour, V. Pavinato, K. R. Andrews, S. Joost, E. L. Landguth, S. Manel, and B. R. Forester. 2017. Effect of missing data and sample size on the performance of genotype-environment association methods.
Yeaman, S., K. A. Hodgins, K. E. Lotterhos, H. Suren, S. Nadeau, J. C. Degner, K. A. Nurkowski, P. Smets, T. Wang, L. K. Gray, K. J. Liepe, A. Hamann, J. A. Holliday, M. C. Whitlock, L. H. Rieseberg, and S. N. Aitken. 2016. Convergent local adaptation to climate in distantly related conifers. Science 353:1431–1433.
Ying, C. C., and Q. Liang. 1994. Geographic pattern of adaptive variation of lodgepole pine (Pinus contorta Dougl.) within the species' coastal range: field performance at age 20 years. Forest Ecology and Management 67:281–298.
Zimin, A., K. A. Stevens, M. W. Crepeau, A. Holtz-Morris, M. Koriabine, G. Marçais, D. Puiu, M. Roberts, J. L. Wegrzyn, P. J. de Jong, D. B. Neale, S. L. Salzberg, J. A. Yorke, and C. H. Langley. 2014. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics 196:875–890.

Table 3.1 Principal component analysis on allele frequencies with a total of 4,155,896 SNPs for 223 individuals of *Pinus ponderosa.*

K	proportion explained variation	*p*-value[a]
PC1	0.7813	8.00e-09 ***
PC2	0.7420	8.00e-09 ***
PC3	0.6934	0.5001

Tracy-Widom test: * = p < 0.05; ** = p < 0.01; *** =p < 0.001.

Table 3.2 SNP annotation with SnpEFF for SNPs significantly associated with Mean maximum temperature of summer (TMAX), April 1st snow pack (PCK4), Mean climatic water deficit (CWD), and Mean minimum temperature of winter (TMIN).

Variant type	TMAX	PCK4	CWD	TMIN
intergenic	172	1505	103	286
downstream	2	29	4	6
intragenic	11	142	4	31
synonymous	5	27	6	5
upstream	12	53	3	5
missense	10	41	0	2
other	1	1	0	0
Total	213	1798	120	335

Table 3.3 Gene ontology for selected environmentally-associated SNPs

SNP	Variant type	Function	Climatic variables
V500239	downstream (UPL6)	Ubiquitination	CWD, TMIN
V2200849	upstream (CKAN _00899300)	Ubiquitination/ deubiquitination	CWD, TMIN
V20216	upstream (LOC101490788)	Ubiquitination	PCK4
V188	intragenic (POPTR_012G114000)	Ubiquitination	PCK4
V236215	intragenic (POPTR_018G017000)	Deubiquitination	PCK4
V4149648	upstream (CISIN_1g0374611mg)	Deubiquitination	TMAX
V1706945	synonymous (DOG1)	ABA-signaling, seed dormancy	CWD
V226436	upstream (PED1)	Regulation of ABA signaling pathway, jasmonic acid biosynthesis, acetyl-CoA C-acyltransferase	PCK4
V827507	upstream (At2g30020)	ABA signaling pathway, fungal defense, wounding response	PCK4
V874448	upstream (ADH1)	Alcohol dehydrogenase; ABA response, abiotic stress responses	PCK4
V2430900	missense (CRRSP38)	ABA response	PCK4
V2860171	missense (ASPG2)	aspartic-type endopeptidase; may be involved in ABA response	PCK4
V2580017	synonymous (Expansis-A8)	Cell wall organization	CWD, TMIN
V977622	upstream (DVH24_020216)	Pectinesterase, cell wall modification	PCK4
V936328	synonymous (CET1)	Reproductive structure initiation	PCK4
V2497819	intragenic (NPY3)	Flower development; gravitropism; ubiquitination	PCK4
V1475638	upstream (JGB)	Negatively regulates pollen germination	PCK4
V252751	downstream (OVA5)	Isoleucine--tRNA ligase	TMAX

		Ovule development; mitochondrial translation	
V3376786	intragenic (CYP94B3)	Anther, stigma, pollen, and fruit development; Jasmonic acid metabolism; wounding and insect defense	TMIN
V544191	missense (VCC)	Vascular tissue histogenesis; cotyledon/leaf development	PCK4
V794961	missense (ERF1A)	Vascular tissue histogenesis; cell division; ethylene-activated signaling pathway; defense	PCK4
V871503	missense (AMTR_s00029p00093020)	Xylem and phloem pattern formation; wounding responses	PCK4
V205892	upstream (AMTR_s00050p00190920)	Uncharacterized protein linked to RNA binding and vascular histogenesis	TMAX
V26357/ V26294	synonymous/upstream (RCN11)	Shoot development, xylosyltransferase activity, involved in seed germination	PCK4
V361321	upstream (PSP)	Embryo, pollen, and root development	PCK4
V2104	intragenic (AMTR_s00086p00155510)	Uncharacterized protein linked to cell division	PCK4
V355233	missense (CYCU1-1)	Cell division	TMAX
V3495148	upstream (DRP3A)	Cell division	TMAX
V2606140	missense (SMAX1)	Seed germination and seedling development	PCK4
V2871460	downstream (BTAF1)	Positive regulation of shoot apical meristem development	PCK4
V2900744	missense (BHLH140)	DNA replication/repair; regulation of secondary shoot formation	TMAX
V1869244	synonymous (CYCA2-3)	Cell division, lateral root formation, stomatal development	TMAX
V3052564	synonymous (SBT1.2)	Cell division, stomatal complex morphogenesis	TMIN
V3160030	synonymous (RPD1)	lateral root morphogenesis, cell division	TMAX
V58925	upstream (GSH2)	Jasmonic acid response (biotic & abiotic stress response)	TMAX

V2423550	missense (FDH)	cuticular wax and suberin biosynthesis; response to cold and light stimulus	PCK4
V1193343	upstream (SELMODRAFT_444240)	Growth regulation and abiotic stress response	PCK4
V3861133	synonymous (4CL)	4-coumarate-CoA ligase, involved in phenylpropanoid pathway (anthocyanins, flavonoids, lignin)	PCK4
V1847064	downstream (4CL1)	4-coumarate-CoA ligase, involved in phenylpropanoid pathway (anthocyanins, flavonoids, lignin)	TMAX
V3946851	upstream (ASD2)	Alpha-L-arabinofuranosidase; Possibly involved in secondary wall formation & leaf abscission	PCK4
V3278803/ V3278802	synonymous/missense (BIG)	Auxin transport protein connected to lateral root formation and fungal pathogen defense	PCK4
V360205	missense (CAMTA1)	Calmodulin-binding transcription activator involved in response to auxin, freezing, & drought	PCK4
V3689405	downstream (CPK29)	Stomatal regulation, fungal pathogen defense, salt tolerance	PCK4
V1961287	synonymous (PBL10)	Stomatal regulation, pathogen defense	PCK4
V1291321	intragenic (LOC104587219)	Regulation of stomata and mitotic cell division	TMIN
V182112	synonymous (HSP70-1)	Heat-shock protein linked to biotic & abiotic stress response	PCK4
V2562111	synonymous (RLK5)	Protein kinase linked to biotic & abiotic stress response	PCK4
V1030420	synonymous (RPA2-6)	Ethylene-responsive transcription factor involved in biotic & abiotic stress response	PCK4
V1197299	synonymous (LOC103718892)	Zinc-ion binding, linked to biotic & abiotic stress	PCK4

		response	
V2371749	synonymous (PCMP-H24)	Zinc-ion binding, linked to biotic & abiotic stress response	PCK4
V1939429	synonymous (PCMP-H43)	Zinc-ion binding, linked to biotic & abiotic stress response	PCK4
V1676572	synonymous (TAO1)	Bacterial pathogen response	PCK4
V2688536	upstream (NQR)	NADPH:quinone oxidoreductase; Bacterial pathogen defense	PCK4
V260166	downstream (AMC1)	Defense/ programmed cell death	PCK4
V2288200	synonymous (LOC107818721)	disease resistance, signal transduction	PCK4
V2968038	intragenic (L484_022527)	Terpenoid (defensive chemical) biosynthesis	PCK4
V2157327	missense (At4g10780)	Defense response	TMIN
V2918743	synonymous (PHO1-3)	Phosphate transport	PCK4
V2269173	upstream (NPF2.13)	Nitrate transport	PCK4
V2881124	downstream (CTI12_AA142310)	PEP carboxylase activity (photosynthesis)	PCK4
V3663893	downstream (CKAN_00949400)	Pyruvate dehydrogenase component (glycolysis)	PCK4
V300383	upstream (AMTR_s00106p00019920)	Sucrose synthesis	PCK4
V3032541	upstream (TPS6)	Trehalose synthesis (can be involved in cold/drought response)	PCK4
V2862091	missense (MST3/STP1)	Glucose import (cell wall formation?)	TMAX
V2433948	upstream (FRO6)	Ferric reduction oxidase; Response to light stimulus	PCK4

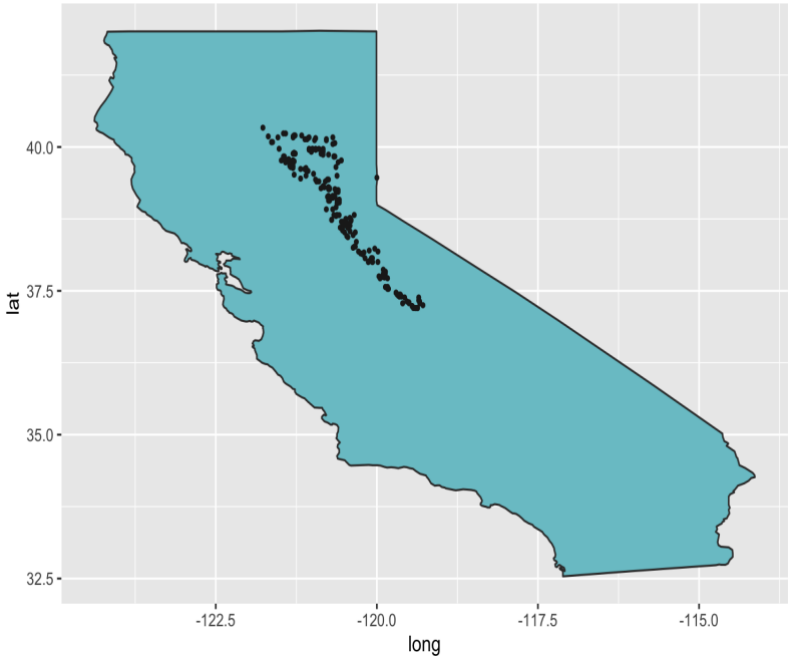

Figure 3.1 Geographic distribution of the 223 ponderosa pine individuals. The black dots represent original genotype source locations.

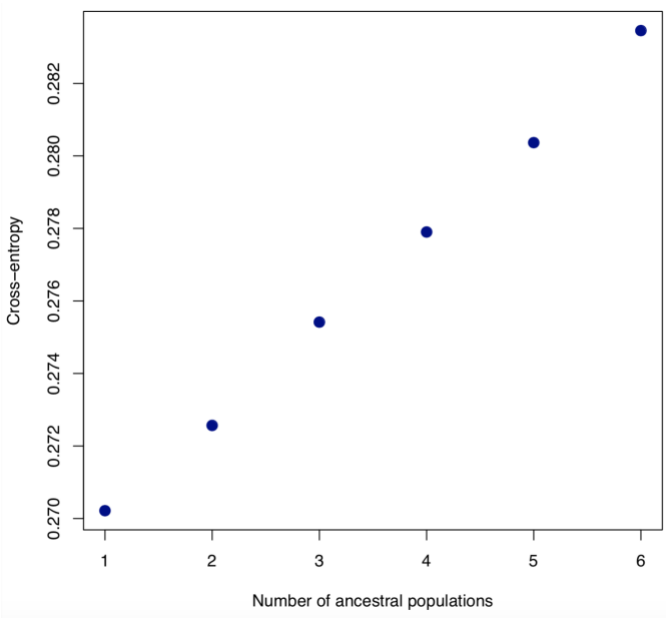

Figure 3.2 Plot of Cross-validation (CV) error of 223 ponderosa pine individuals based on a total of 4,155,896 SNPs at K=1, 2, 3, 4, 5, 6. K represents the number of populations.

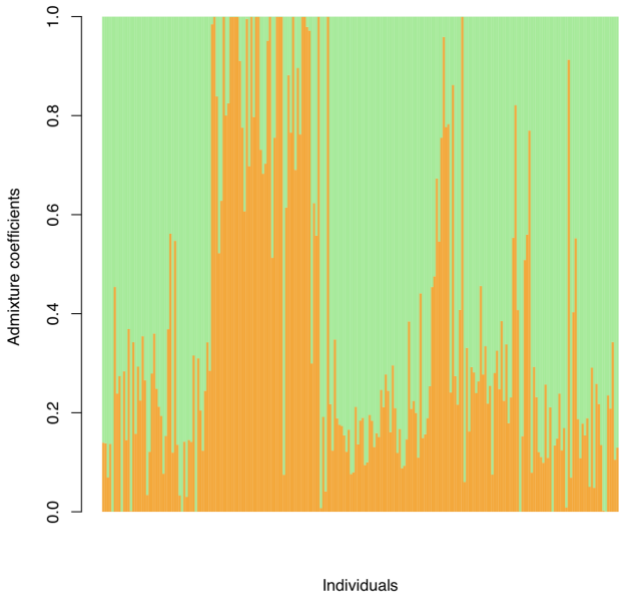

Figure 3.3 Admixture analysis of 223 individuals based on a total of 4,155,896 SNPs at K=2. K represents the number of populations.

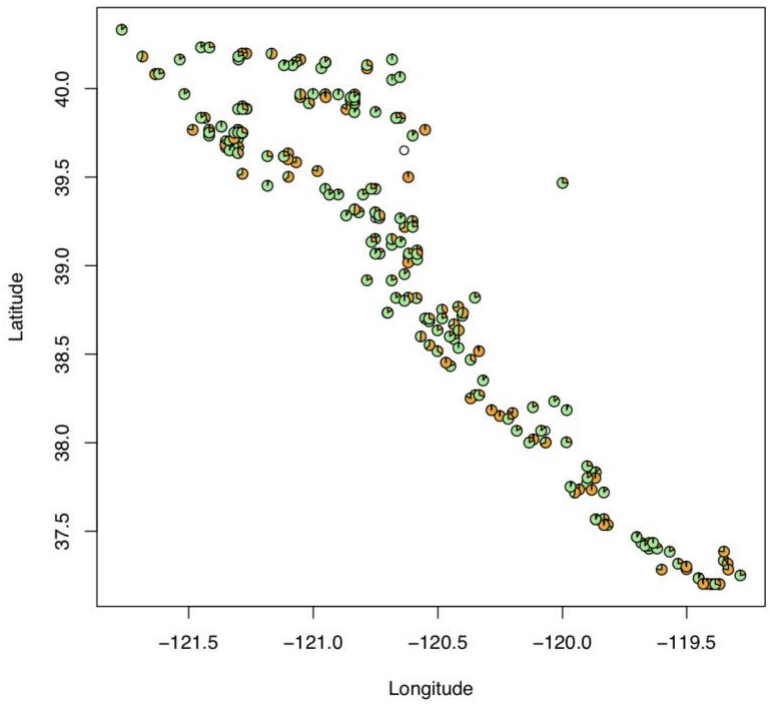

Figure 3.4 Population structure of the 223 individuals based on a total of 4,155,896 SNPs. Genetic assignments under K = 2 based on admixture results. K represents the number of populations. The circle represents the location of each individual. The proportion of yellow and green in the circle represents genetic contribution of each of the two populations to each individual.

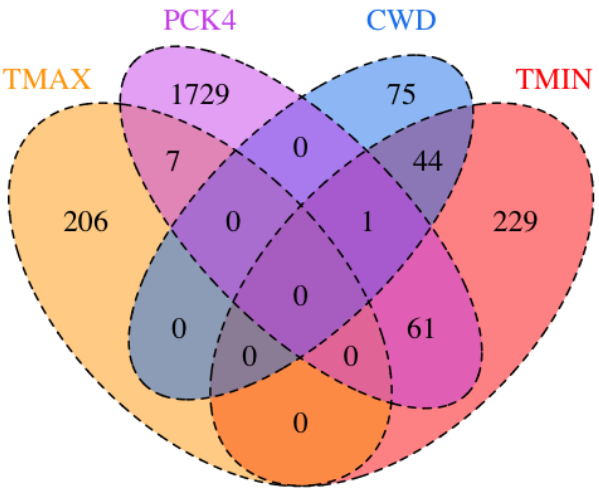

Figure 3.5 Venn diagram comparing SNPs overlap between the ones significantly ($q \leq 0.05$) associated with Mean maximum temperature of summer (TMAX), April 1st snow pack (PCK4), Mean climatic water deficit (CWD), and Mean minimum temperature of winter (TMIN).

Chapter 4
Seedling drought response physiology associated with genetic variation in ponderosa pine

4.0 Abstract

Drought stress is a major cause of tree mortality in Mediterranean coniferous forest in western United States. This study dissects the genetics of drought tolerance traits in ponderosa pine (*Pinus ponderosa*) by combining genotype-to-phenotype (G2P) association analysis with a greenhouse experiment. We collected seeds from 48 genotyped mother trees and planted seeds from each maternal family into 10 tubular pots in both dry and wet water treatments. Eight phenotypic traits were measured during or after the greenhouse experiment. Six drought-responsive traits were identified, including RL (root length), GR (height growth), SW (shoot weight), R2S (root-shoot dry mass ratio), SDAD (stomata density on adaxial side), NRAB (number of stomatal rows on abaxial side). Seedlings exposed to drought exhibit larger RL, R2S, SDAD, NRAB, and lower GR and SW. We ran a G2P analysis using maternal genotype (chapter 3) and the breeding values for these 6 traits using LFMM2. We found 153 SNPs strongly associated with RL, 80 with SW, 145 with GR, 42 with SDAD, 85 with NRAB, and 1530 with R2S. The identified SNPs reside in or near genes with a wide variety of functions, including ubiquitination, abscisic acid (ABA) signaling pathway, cell division or growth of roots or shoots, and cell wall organization. Roots play a critical role in both the drought responsive traits and the function of correlated genes, which need to be incorporated to understand the response of pine trees to climate change in future studies.

4.1 Introduction

Under the on-going anthropogenic climate change, longer, more frequent and more intense drought periods are predicted in California, which already has a summer-dry Mediterranean climate (Giorgi and Lionello 2008, IPCC 2014). Water stress will therefore be a leading constrain on plant survival and productivity and has already been driving drought-induced mortality in California and other dry forests around the world (Loarie et al. 2009, Pereira et al. 2010, Allen et al. 2010). The potential to adapt to the new environmental conditions can be achieved by phenotypic plasticity at the individual level, and either genetic adaptation or range shift at the population level (Aitken et al. 2008, Anderson et al. 2012, Alberto et al. 2013). A better understanding of adaptive genetic variation can help clarify to what extent Mediterranean trees are adapted to current moisture gradients and can help us better predict and manage forests in a changing climate (White et al. 2007, Neale and Kremer 2011). In particular, the first-year seedling stage is a bottleneck for the establishment and growth of forest species because seedlings are highly susceptible to resource limitations and have much higher mortality than established individuals (Grubb 1977, Leck et al. 2008).

Some traits thought to be involved in drought tolerance of conifer trees include root-to-shoot ratio, root biomass and length, specific leaf area (SLA, the ratio of leaf area to dry mass), stomatal conductance, and water use efficiency (WUE, the ratio of CO_2 assimilation to transpiration) (Picon et al. 1996, Cregg and Zhang 2001, de Miguel et al.

2012, Olmo et al. 2014, Moran et al. 2017). Most recent research on drought stress has focused on aboveground tree parts (McDowell et al. 2008, Ryan 2011, Hamanishi and Campbell 2011) due to the difficulties in observing and studying roots (Brunner et al. 2015). However, roots play an important role in drought responses of trees by being responsible for water uptake and acting as sensors for water-deficit conditions (Brunner and Godbold 2007, Hamanishi and Campbell 2011). For example, studies show that seedlings of trees also increase allocation of biomass to roots to augment water acquisition during drought (Markesteijn and Poorter 2009). Thus, a better understanding of the physiology and genetics of root traits would improve our understanding of drought tolerance in forest trees.

Multiple provenance studies have shown that genetic differences are likely to play an important role in drought-tolerance (McDowell et al. 2008, Moran et al. 2017). For example, seedlings from dry environments often exhibit more root growth, higher drought survival (Cregg and Zhang 2001), and higher WUE (Cregg et al. 2000, Voltas et al. 2008). Moreover, several studies have investigated changes in gene expression in drought-stressed conifer seedlings (Ralph et al. 2006, Hamanishi and Campbell 2011). Some of the genes identified include late-embryogenesis-abundant (LEA) proteins - involved in seed dormancy, which also requires tolerance of dry conditions - and abscissic acid (ABA) signaling pathways, which are involved in stomatal regulation. (Ralph et al. 2006, Hamanishi and Campbell 2011). Most of the gene expression changes return to normal after re-watering the drought-stressed seedlings; such changes are likely responsible for plastic environmental responses, rather than locally adaptive differences in mean trait values (Bräutigam et al. 2013).

Some fundamental questions are still largely unresolved, including the nature and number of genes involved in drought-tolerance traits (Barton and Keightley 2002, Prunier et al. 2011). Recently, there has been an increase in the use of genetic association techniques to identify genes underlying quantitative traits in forest trees (Eckert et al. 2010a). A quantitative trait is a phenotype that exhibits continuous variation due to the cumulative actions of many genes. Genotype-to-Phenotype (G2P) association, which identifies loci linked to a particular quantitative trait by correlating genotypes at SNPs with the variation in certain traits (Eckert et al. 2009, Holliday et al. 2010), can help to identify the genes underlying the drought response traits of trees. To reduce or eliminate the effects of environmental differences on phenotypes, traits must be measured in a common environment, such as a greenhouse or common garden.

Ponderosa pine (*Pinus ponderosa*) is a highly valuable Mediterranean coniferous species in the western United States, where it is a widely adapted and ubiquitous conifer (Conkle and Critchfield 1988). However, the genetics of drought tolerance traits in this species are largely unknown. Some studies have emphasized the importance of intraspecific variation of *P. ponderosa* for responses to climate change, but mainly focus on the phenotypic variation within and among populations rather than the genetic basis of this variation (Kitzmiller 2005, Kolb et al. 2016). A deep understanding of the genetic basis of drought tolerance traits in ponderosa pine is critical for successful reforestation, for conservation and restoration programs, and for potentially coping with climate-induced species range changes.

The main objective of this work was to unravel the genetic basis of different physiological traits in response to drought for *P. ponderosa* seedlings. For this purpose, a greenhouse experiment together with a G2P association analysis was employed. We used SNPs derived from GBS of widely distributed individuals of ponderosa pine from the Sierra Nevada. Our specific research objectives were the identification of: 1) physiological traits in response to drought in the greenhouse experiment with 48 maternal families of seedlings grown under both dry and wet conditions; 2) identification of loci and genomic regions underlying these traits through G2P analysis; and 3) identification of a set of promising candidate genes and the annotated function involved with these loci.

4.2 Materials and method

4.2.1 Common garden procedure

In the 1970's, the Forest Service's Pacific Southwest Regional Genetic Resources Program planted clones of 302 wild ponderosa pines in Chico, California. They came from diverse climate conditions in the central portion of California's Sierra Nevada mountains (Chapter 1). For this study, wind-pollinated seed was collected from 50 parent trees of *P. ponderosa* among those already genotyped (Chapter 3). For each family, I collected two to three cones during summer 2018 and put them into paper bags. Because pines are wind-pollinated outcrossing species (Williams 2009), seeds from the same tree are mostly half-siblings, occasionally full-sibs.

Once the cones were dry, I collected all the seeds from each individual and stored them in the refrigerator. During winter 2018, the collected seeds were stratified to break dormancy. Seeds were place in an open tank with water, which was shaken several times to aerate the water, for 48 hours. After soaking, seeds were surface-dried and then placed in plastic bags in the refrigerator (~1.7°C) for 6 weeks. Only 48 families had enough seeds for the greenhouse experiment.

We aimed to have 10 seedlings from each maternal family in both wet and dry treatments, 1000 seedlings in total. We used plastic tubes 8-cm in diameter and 120-cm long for planting, because the maximum root length in a pilot experiment conducted in 2017 was more than 110 cm. We used PVC pipes to also build 10 frames with 100 tubes in each frame. The planting soil was a mixture of 70% sand, 20% vermiculite, and 10% organic-rich potting mix to mimic the coarse texture of the soil of Sierra Nevada conifer forests (Bales et al. 2011).

Seeds were planted in February 2019. Two seeds from each family were planted into each tube to allow for failed germination, with two tubes from each family per frame (20 seedlings per family total). Tubes were labeled with family ID and randomly placed within each frame. In April 2019, we replanted more seeds in any tubes without seedlings. All the tubes were watered every other day during the germination and seedling establishment period (February through June).

At the end of June 2019, extra seedlings were removed, and alternating frames were assigned to the wet treatment and the dry treatment (five frames containing up to 500 seedlings per treatment) (Figure 4.1). The wet treatment group was watered twice every week and the drought treatment group was watered once every three weeks until mid-October (3.5 months). While wild ponderosa pine seedlings would receive little to no precipitation during the summer months, this occasional watering was necessary in the

greenhouse environment even in the dry treatment to prevent complete mortality. Temperatures inside the greenhouse in the low-elevation environment of Merced, CA could reach as high as 37 °C on the hottest days and the soil volume of the tubes was limited, with no access to groundwater, both of which could make evaporation and drought stress more intense than the no-precipitation condition in the wild.

4.2.2 Phenotypic measurements and analysis

Several phenotypic traits were measured during and after the greenhouse experiment. Before implementing the drought treatment, we recorded the initial shoot height of each seedling. We also measured the final height of the seedlings before they were harvested in October. Thus, we calculated shoot growth as final height minus initial height. Immediately after the harvesting (to avoid shrinkage), the length of fresh roots was measured from soil surface to taproot tip. Following harvest, needles, fresh stem and fresh roots of all the seedlings were separately put into paper lunch bags and dried in the oven at 75 °C for 48 hours. We measured root dry mass (RW) as well as shoot weight (SW, total of stem and needles). We then calculated root-shoot ratio (R2S) as RW/SW. Specific Root length (SRL) was calculated as root length/root weight.

Before harvest, we also collected 3-4 fresh needles for each living seedling to calculate stomatal density. In pines, stomata are arranged into longitudinal rows. We put each needle on a slide and photographed it at 100x magnification using a Leica DME compound microscope equipped with a Leica DFC290 digital camera. All counts were conducted near the middle of the needle to avoid variation that might occur at the base and at the tip. Approximately 1.96 mm lengths of needle randomly placed along its adaxial (upper) and abaxial (lower) surfaces were surveyed for number of stomata and stomatal rows. Needle width was measured in magnified images using the line measure tool in the Leica software. Then we calculated the stomata density on each side as the number of stomata divided by 1.96 times needle width. Individual seedling means were calculated by averaging abaxial (AB) stomatal density of AB and number of stomatal rows on both sides (AB & AD) across sampled needles.

In summary, the following eight traits were recorded in the greenhouse experiment: height growth (GR; cm), root length (RL; cm), dry shoot weight (SW; g), dry root weight (RW; g), the ratio of root to shoot dry mass (R2S), specific root length (SRL; cm/g), stomata density of adaxial side (SDAB; /mm^2), number of stomatal rows on abaxial side (NRAB), and the number of stomatal rows on abaxial side (NRAD). Only 42 out of 48 mother trees had enough germination to carry out these measurements across both treatments. After obtaining the these phenotypic data, we ran analyses of variance (ANOVA) in R (www.rproject.org) to test which phenotypic traits are significantly related to drought treatment by comparing the data from wet and dry treatment while accounting for block (planting box-level) differences.

4.2.3 Genotype-phenotype association analysis

The 42 individual mother trees had already been genotyped for over 4 million SNPs using GBS (Chapter 3). We used these same SNPs for the G2P association analysis, focusing on the traits significantly associated with drought treatments. The breeding value (BV) of a tree for a given trait (that is, biomass production or height) accounts for

the sum of gene effects that contribute to that trait. An individual with a high BV for root length, for instance, tends to produce offspring with long roots. The BV of an individual is estimated by measuring the relatives (Meuwissen et al. 2001, Isik 2014), in this case as the average trait value for the 10 offspring in the wet treatment. We used LFMM 2 (Caye et al. 2019) (details in Chapter 3) to run the genotype to phenotype association analysis, and then identified associations based on p ($<10^{-5}$) value.

4.2.4 Gene annotation

After we identified the significantly associated SNPs, we ran SnpEff (Cingolani et al. 2012) for SNP annotation. We built the data base with the annotated genome and the reference genome of loblolly pine v.2.01 in TreeGenes (http://treegenesdb.org/FTP/Genomes/Pita/v2.01/). Then we aligned the gene sequence against the nonredundant protein sequences database using UniProt to identify the gene and protein with the implemented Blastx (2.9.0+, e < 1e−10). The Gene Ontology Annotation Database ("UniProt" 2015, Bateman et al. 2017) was used to further identify the potential functions of the genes.

4.3 Results

4.3.1 Drought responsive traits

As shown in Table 4.1 and Figure 4.1, six out of the eight measured phenotypic traits were significantly different in the drought treatment versus the wet treatment, including RL (root length), GR (height growth), SW (shoot weight), R2S (root-shoot dry mass ratio), and SDAD (stomatal density on adaxial side), NRAB (number of stomatal rows on abaxial side). In the drought treatment, seedlings grew longer roots, gained less dry shoot height and mass, had higher stomatal density on adaxial side of the needle and more stomatal rows on abaxial side, and had a higher root to shoot dry mass ratio (Figure 4.2).

4.3.2 Phenotypic associations at individual loci

After the running of LFMM2, we found many significant associations between SNPs and the six phenotypic variables. There are 153 SNPs strongly associated with root length, 80 with shoot weight, 145 with height growth, 42 with stomata density on adaxial side, 85 with number of stomatal rows on abaxial side, and 1530 with root-shoot ratio. Only a few SNPs overlapped among these six traits. There are 21 SNPs associated with both root length and root-shoot ratio, and 26 SNPs associated with both root-shoot ratio and height growth, but no overlapping among these two sets of SNPs.

4.3.3 Gene annotation

Table 4.2 shows locations of SNPs relative to genes according to SnpEFF. Catagories intragenic (intron) variants, intergenic variants, upstream SNPs, downstream SNPs, and synonymous or missense variants in the gene coding sequence. According to the user manual of SnpEFF, missense variant is defined as non-synonymous variant. Most of the SNPs are between genes (in the intergenic regions) and likely have no direct effect on gene expression. For the gene annotation, we focused on the other five types of SNP variant.

Table 4.3 shows selected phenotypically associated SNPs. We found several protein types that are likely relevant to the drought responses. Some of the SNPs associated with RL, GR, R2S and SDAD are in or near genes in the protein ubiquitination pathway. Some SNPs associated with RL are linked to the jasmonic acid synthesis pathway. NRAB (number of stomatal rows of abaxial side) and R2S were associated with SNPs in the abscissic acid (ABA) signaling pathway. The ABA pathway is involved in stomatal closure in response to drought stress, while protein ubiquitination and jasmonic acid signaling are involved in both biotic and abiotic stress responses (Moran et al. 2017). Only GR was associated with SNPs in near genes involved in seed dormancy, which can also be involved in drought responses (Moran et al. 2017). NRAB and R2S were also associated with genes involved in cell wall organization and pectin synthesis. Genes involved in auxin biosynthesis in roots, root hair and lateral root formation, were associated with R2S, SW (shoot weight), and NRAB. All of these processes are related to plant cell division and root and shoot growth. GR was associated with genes involved in stomatal development and cell division.

For many of the other loci associated with phenotypic variables, gene ontology results were too vague to draw many conclusions about their function or why the association might exist. However, some of these genes have been previously associated with plant stress, including leucine-rich protein with stress responses in roots (Park et al. 2014), Inosine-uridine preferring nucleoside hydrolase with drought responses in roots (Micheletto et al. 2007), pentatricopeptide repeat-containing protein to abiotic and biotic stress (Xing et al. 2018), Metallophos protein to drought (Gugger et al. 2017), and Retrovirus-related Pol polyprotein from transposon TNT 1-9 related to biotic and abiotic stress (Huang et al. 2018). Most of the others are involved in gene expression (RNA or DNA binding, mRNA process, helicase activity, ribosome components, methylation) or ATP binding.

4.4 Discussion

Six drought-response-related traits were found in this study, including RL (root length), SW (shoot weight), R2S (root-shoot dry mass ratio), SDAD (stomata density on adaxial side), NRAB (number of stomatal rows on abaxial side), and GR (growth). Our study shows that drought-stressed ponderosa pine seedlings allocate more investment to their root system than to shoots, with longer root length, higher root to shoot dry mass ratio, less dry shoot mass and less height growth. Other studies have found a similar pattern in pines. For example Taeger et al. (2015) and Cregg and Zhang (2001) both identified a plastic response to drought by increased taproot length and root–shoot ratios in Scots pine (*Pinus sylvestris*) seedlings. Root growth of loblolly pine (*Pinus taed* seedlings was affected more than shoot growth by water stress, causing a higher root-shoot ratio (Seiler and Johnson 1988, Cregg and Zhang 2001). Shoot and needle growth were significantly reduced in drought-treatment in *P. sylvestris* (Irvine et al. 1998). This may indicate acclimation to at the cost of overall low growth of aboveground structures in pines in response to dry soil.

A positive relationship was identified between drought and both stomata density on the adaxial side and number of stomatal rows on abaxial side. Studies of different tree species have yielded conflicting results concerning how stomatal traits respond to

drought stress. Some studies have showed evidences of increasing stomatal density and a reduction in stomata size as a response to drought, interpreted as an adaptive response allowing for more sensitive stomatal regulation (Dunlap and Stettler 2001, Pearce et al. 2006). However, other studies showed contradictory results. For example, Schoettle and Rochelle (2000) identified a significant decrease in stomatal density in limber pine (*Pinus flexilis*) with increasing elevation. Higher elevations in this case were drier than lower ones, suggesting that changes may be related to conserving water. In our study, a high stomatal density and number of stomatal rows, which can lead to a higher leaf gas interchange in short favorable periods and more control of water loss and gas exchange under drought stress (Afas et al. 2007), may be advantageous for ponderosa pine seedlings growing in harsh dry environments. Other studies have found a higher stomatal density and/or number of stomatal rows under drought in other Mediterranean pines, such as *P. canariensis* (López et al. 2008), *P. brutia* (Dangasuk and Panetsos 2004) and *P. pinaster* (Wahid et al. 2006).

Most (1530) of the environmentally-associated SNPs were linked to R2S while only 80 were linked to SW, which may reflect the importance of biomass allocation to roots instead of shoots in response to drought. Consistently, some of the SNPs associated with R2S were in or near genes responsible for root hair formation, lateral root development, drought response in roots, and ABA signaling. The upstream SNP associated with NRAB was also linked to lateral root formation. Though no overlapping SNPs were found between R2S and SW, an overlapping function (auxin biosynthesis in roots) of related genes was found. Studies have identified the critical role of auxin in root development and ABA signaling under water stress, mostly in model species. For example, Xu et al. (2013) found that moderate water stress increased ABA accumulation and auxin transport in the root apex in rice and *Arabidopsis thaliana* plants, which enhanced proton secretion for maintaining primary root elongation and root hair development. Moreover, auxin also positively modulated the lateral root number and ABA-responsive genes expression in *Arabidopsis* under drought stress condition (Shi et al. 2014). Many plants use ABA concentrations as a signal to keep stomata closed (Brodribb et al. 2014) and affect shoot growth and water uptake (Buckley 2005, Hamanishi and Campbell 2011), thus affect the drought response. However, how auxin regulates roots and ABA is largely unknown in conifers. Our study may indicate the critical role of auxin and its regulation on roots and ABA in pines trees under drought condition.

The prevalence of genetic associations related to ubiquitination, including SNPs associated with RL, GR, R2S and SDAD is consistent with prior studies of drought response in conifers (Moran et al. 2017). Ubiquitin has been found to be involved in drought responses in model species by playing a role in ABA-mediated dehydration stress responses (Ryu et al. 2010) or by the downregulation of plasma membrane aquaporin levels (Lee et al. 2009). For example, Kim et al. (2012) found that ubiquitin protein was induced in developing lateral roots, root tips and the vascular tissues of tap roots in ABA-mediated dehydration stress responses in *Arabidopsis*. However, the study of ubiquitin in relation to the drought response in conifer species is limited. One study in black spruce (*Picea mariana*) identified 16 out 313 candidate genes correlated with precipitation, including the genes in the ubiquitin protein handling pathway, but the

related traits were unknown (Prunier et al. 2011). The association between ubiquitin protein and roots and stomatal density may indicate the potential role of ubiquitin protein in roots and stomata responses in conifer species under water stress.

Quite a few genes involved with cell wall organization and pectin catabolic processes were found to be associated with R2S and NRAB. Studies have identified the critical role of cell wall components, such as pectin, in stress response in different plant species (Jarvis 2009, Seifert and Blaukopf 2010, Wolf and Greiner 2012, Le Gall et al. 2015). In conifers, Pattathil et al. (2016) found that the loosening of cell wall pectic components might trigger stress-signaling responses and thus initiate a cascade of stress-mitigating processes in loblolly pine. Another study in *Pinus radiata* found that increases in cell wall elasticity may be related to drought tolerance (De Diego et al. 2013). In addition, the genes involved in cell wall organization might also be involved in xylem formation, as the degree of cell expansion during xylem formation can have a strong effect on the drought resistance of that xylem (Anfodillo et al. 2012, Bryukhanova and Fonti 2013). However, how the cell wall organization and pectin catabolic process differs between organs under drought conditions in trees is still largely unknown. Our study may indicate that changes cell wall composition and pectin catabolic process in response to drought are different in shoots and roots.

To conclude, our study represents a first step in understanding and dissecting the genetic architecture of drought responsive traits in ponderosa pine using G2P association analysis combined with a greenhouse experiment. We identified a total of 1839 SNPs associated with 6 drought responsive traits. The identified SNPs locate in genes with a wide variety of functions. Potentially, the identified genes and alleles are valuable resources for pine breeding through marker assisted selection and genomic selection, specifically under the rapid changing climate scenarios. In addition, roots play a critical role in both the drought responsive traits and the function of correlated genes in our study. Future studies may need to incorporate the roots traits to understand the response of pine trees to changing climate.

4.5 References

Afas, N. A., N. Marron, and R. Ceulemans. 2007. Variability in Populus leaf anatomy and morphology in relation to canopy position, biomass production, and varietal taxon. Annals of Forest Science 64:521–532.

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. Evolutionary Applications 1:95–111.

Alberto, F. J., S. N. Aitken, R. Alía, S. C. González-Martínez, H. Hänninen, A. Kremer, F. Lefèvre, T. Lenormand, S. Yeaman, R. Whetten, and O. Savolainen. 2013. Potential for evolutionary responses to climate change – evidence from tree populations. Global Change Biology 19:1645–1661.

Allen, C. D., A. K. Macalady, H. Chenchouni, D. Bachelet, N. McDowell, M. Vennetier, T. Kitzberger, A. Rigling, D. D. Breshears, E. H. (Ted) Hogg, P. Gonzalez, R. Fensham, Z. Zhang, J. Castro, N. Demidova, J.-H. Lim, G. Allard, S. W. Running, A. Semerci, and N. Cobb. 2010. A global overview of drought and heat-induced

tree mortality reveals emerging climate change risks for forests. Forest Ecology and Management 259:660–684.

Anderson, J. T., A. M. Panetta, and T. Mitchell-Olds. 2012. Evolutionary and Ecological Responses to Anthropogenic Climate Change: Update on Anthropogenic Climate Change. Plant Physiology 160:1728–1740.

Anfodillo, T., A. Deslauriers, R. Menardi, L. Tedoldi, G. Petit, and S. Rossi. 2012. Widening of xylem conduits in a conifer tree depends on the longer time of cell expansion downwards along the stem. Journal of Experimental Botany 63:837–845.

Bales, R. C., J. W. Hopmans, A. T. O'Geen, M. Meadows, P. C. Hartsough, P. Kirchner, C. T. Hunsaker, and D. Beaudette. 2011. Soil Moisture Response to Snowmelt and Rainfall in a Sierra Nevada Mixed-Conifer Forest. Vadose Zone Journal 10:786–799.

Barton, N. H., and P. D. Keightley. 2002. Understanding quantitative genetic variation. Nature Reviews Genetics 3:11–21.

Bateman, A., M. J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-A-Jee, A. Cowley, A. D. Silva, M. D. Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, A. Renaux, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.-C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.-L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L.-S. Yeh, and J. Zhang. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Research 45:D158–D169.

Bräutigam, K., K. J. Vining, C. Lafon-Placette, C. G. Fossdal, M. Mirouze, J. G. Marcos, S. Fluch, M. F. Fraga, M. Á. Guevara, D. Abarca, Ø. Johnsen, S. Maury, S. H. Strauss, M. M. Campbell, A. Rohde, C. Díaz-Sala, and M.-T. Cervera. 2013. Epigenetic regulation of adaptive responses of forest tree species to the environment. Ecology and Evolution 3:399–415.

Brodribb, T. J., S. A. M. McAdam, G. J. Jordan, and S. C. V. Martins. 2014. Conifer species adapt to low-rainfall climates by following one of two divergent pathways. Proceedings of the National Academy of Sciences 111:14489–14493.

Brunner, I., and D. L. Godbold. 2007. Tree roots in a changing world. Journal of Forest Research 12:78–82.
Brunner, I., C. Herzog, M. A. Dawes, M. Arend, and C. Sperisen. 2015. How tree roots respond to drought. Frontiers in Plant Science 6.
Bryukhanova, M., and P. Fonti. 2013. Xylem plasticity allows rapid hydraulic adjustment to annual climatic variability. Trees 27:485–496.
Buckley, T. N. 2005. The control of stomata by water balance. New Phytologist 168:275–292.
Caye, K., B. Jumentier, J. Lepeule, and O. François. 2019. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Molecular Biology and Evolution 36:852–860.
Cingolani, P., A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6:80–92.
Conkle, M. T., and W. B. Critchfield. 1988. Genetic variation and hybridization of ponderosa pine. In: Ponderosa Pine: the species and its management, Washington State University Cooperative Extension, 1988: p. 27-43.
Cregg, B. M., J. M. Olivas-García, and T. C. Hennessey. 2000. Provenance variation in carbon isotope discrimination of mature ponderosa pine trees at two locations in the Great Plains. Canadian Journal of Forest Research 30:428–439.
Cregg, B. M., and J. W. Zhang. 2001. Physiology and morphology of Pinus sylvestris seedlings from diverse sources under cyclic drought stress. Forest Ecology and Management 154:131–139.
Dangasuk, O. G., and K. P. Panetsos. 2004. Altitudinal and longitudinal variations in Pinus brutia (Ten.) of Crete Island, Greece: some needle, cone and seed traits under natural habitats. New Forests 27:269–284.
De Diego, N., M. C. Sampedro, R. J. Barrio, I. Saiz-Fernández, P. Moncaleán, and M. Lacuesta. 2013. Solute accumulation and elastic modulus changes in six radiata pine breeds exposed to drought. Tree Physiology 33:69–80.
Dunlap, J. M., and R. F. Stettler. 2001. Variation in leaf epidermal and stomatal traits of Populus trichocarpa from two transects across the Washington Cascades. Canadian Journal of Botany 79:528–536.
Giorgi, F., and P. Lionello. 2008. Climate change projections for the Mediterranean region. Global and Planetary Change 63:90–104.
Grubb, P. J. 1977. The Maintenance of Species-Richness in Plant Communities: The Importance of the Regeneration Niche. Biological Reviews 52:107–145.
Gugger, P. F., J. M. Peñaloza-Ramírez, J. W. Wright, and V. L. Sork. 2017. Whole-transcriptome response to water stress in a California endemic oak, Quercus lobata. Tree Physiology 37:632–644.
Hamanishi, E. T., and M. M. Campbell. 2011. Genome-wide responses to drought in forest trees. Forestry: An International Journal of Forest Research 84:273–283.
Huang, B.-H., Y.-C. Lin, C.-W. Huang, H.-P. Lu, M.-X. Luo, and P.-C. Liao. 2018. Differential genetic responses to the stress revealed the mutation-order adaptive divergence between two sympatric ginger species. BMC Genomics 19:692.

IPCC. 2014. Climate Change 2014: Impacts,Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Page 1132. Cambridge University Press, Cambridge, UK and New York, USA.

Irvine, J., M. P. Perks, F. Magnani, and J. Grace. 1998. The response of Pinus sylvestris to drought: stomatal control of transpiration and hydraulic conductance. Tree Physiology 18:393–402.

Isik, F. 2014. Genomic selection in forest tree breeding: the concept and an outlook to the future. New Forests 45:379–401.

Jarvis, M. C. 2009. Plant cell walls: supramolecular assembly, signalling and stress. Structural Chemistry 20:245–253.

Kim, S. J., M. Y. Ryu, and W. T. Kim. 2012. Suppression of Arabidopsis RING-DUF1117 E3 ubiquitin ligases, AtRDUF1 and AtRDUF2, reduces tolerance to ABA-mediated drought stress. Biochemical and Biophysical Research Communications 420:141–147.

Kitzmiller, J. H. 2005. Provenance Trials of Ponderosa Pine in Northern California. Forest Science 51:595–607.

Kolb, T. E., K. C. Grady, M. P. McEttrick, and A. Herrero. 2016. Local-Scale Drought Adaptation of Ponderosa Pine Seedlings at Habitat Ecotones. Forest Science 62:641–651.

Le Gall, H., F. Philippe, J.-M. Domon, F. Gillet, J. Pelloux, and C. Rayon. 2015. Cell Wall Metabolism in Response to Abiotic Stress. Plants 4:112–166.

Leck, M. A., V. T. Parker, R. L. Simpson, and R. S. Simpson. 2008. Seedling Ecology and Evolution. Cambridge University Press.

Lee, H. K., S. K. Cho, O. Son, Z. Xu, I. Hwang, and W. T. Kim. 2009. Drought Stress-Induced Rma1H1, a RING Membrane-Anchor E3 Ubiquitin Ligase Homolog, Regulates Aquaporin Levels via Ubiquitination in Transgenic Arabidopsis Plants. The Plant Cell 21:622–641.

Loarie, S. R., P. B. Duffy, H. Hamilton, G. P. Asner, C. B. Field, and D. D. Ackerly. 2009. The velocity of climate change. Nature 462:1052–1055.

López, R., J. Climent, and L. Gil. 2008. From desert to cloud forest: the non-trivial phenotypic variation of Canary Island pine needles. Trees 22:843.

Markesteijn, L., and L. Poorter. 2009. Seedling root morphology and biomass allocation of 62 tropical tree species in relation to drought- and shade-tolerance. Journal of Ecology 97:311–325.

McDowell, N., W. T. Pockman, C. D. Allen, D. D. Breshears, N. Cobb, T. Kolb, J. Plaut, J. Sperry, A. West, D. G. Williams, and E. A. Yepez. 2008. Mechanisms of plant survival and mortality during drought: why do some plants survive while others succumb to drought? New Phytologist 178:719–739.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819–1829.

Micheletto, S., L. Rodriguez-Uribe, R. Hernandez, R. D. Richins, J. Curry, and M. A. O'Connell. 2007. Comparative transcript profiling in roots of Phaseolus acutifolius and P. vulgaris under water deficit stress. Plant Science 173:510–520.

de Miguel, M., D. Sanchez-Gomez, M. T. Cervera, and I. Aranda. 2012. Functional and genetic characterization of gas exchange and intrinsic water use efficiency in a full-sib family of *Pinus pinaster* Ait. in response to drought. Tree Physiology 32:94–103.

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

Neale, D. B., and A. Kremer. 2011. Forest tree genomics: growing resources and applications. Nature Reviews Genetics 12:111–122.

Olmo, M., B. Lopez-Iglesias, and R. Villar. 2014. Drought changes the structure and elemental composition of very fine roots in seedlings of ten woody tree species. Implications for a drier climate. Plant and Soil 384:113–129.

Park, S., J.-C. Moon, Y. C. Park, J.-H. Kim, D. S. Kim, and C. S. Jang. 2014. Molecular dissection of the response of a rice leucine-rich repeat receptor-like kinase (LRR-RLK) gene to abiotic stresses. Journal of Plant Physiology 171:1645–1653.

Pattathil, S., M. W. Ingwers, O. L. Victoriano, S. Kandemkavil, M. A. McGuire, R. O. Teskey, and D. P. Aubrey. 2016. Cell Wall Ultrastructure of Stem Wood, Roots, and Needles of a Conifer Varies in Response to Moisture Availability. Frontiers in Plant Science 7.

Pearce, D. W., S. Millard, D. F. Bray, and S. B. Rood. 2006. Stomatal characteristics of riparian poplar species in a semi-arid environment. Tree Physiology 26:211–218.

Pereira, H. M., P. W. Leadley, V. Proença, R. Alkemade, J. P. Scharlemann, J. F. Fernandez-Manjarrés, M. B. Araújo, P. Balvanera, R. Biggs, and W. W. Cheung. 2010. Scenarios for global biodiversity in the 21st century. Science 330:1496–1501.

Picon, C., J. M. Guehl, and A. Ferhi. 1996. Leaf gas exchange and carbon isotope composition responses to drought in a drought-avoiding (Pinus pinaster) and a drought-tolerant (Quercus petraea) species under present and elevated atmospheric CO2 concentrations. Plant, Cell & Environment 19:182–190.

Prunier, J., J. Laroche, J. Beaulieu, and J. Bousquet. 2011. Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce: SNPs and climate adaptation. Molecular Ecology 20:1702–1716.

Ralph, S. G., H. Yueh, M. Friedmann, D. Aeschliman, J. A. Zeznik, C. C. Nelson, Y. S. Butterfield, R. Kirkpatrick, J. Liu, and S. J. Jones. 2006. Conifer defence against insects: microarray gene expression profiling of Sitka spruce (Picea sitchensis) induced by mechanical wounding or feeding by spruce budworms (Choristoneura occidentalis) or white pine weevils (Pissodes strobi) reveals large-scale changes of the host transcriptome. Plant, Cell & Environment 29:1545–1570.

Ryan, M. G. 2011. Tree responses to drought. Tree Physiology 31:237–239.

Ryu, M. Y., S. K. Cho, and W. T. Kim. 2010. The Arabidopsis C3H2C3-Type RING E3 Ubiquitin Ligase AtAIRP1 Is a Positive Regulator of an Abscisic Acid-Dependent Response to Drought Stress. Plant Physiology 154:1983–1997.

Schoettle, A., and S. G. Rochelle. 2000. Morphological variation of Pinus flexilis (Pinaceae), a bird-dispersed pine, across a range of elevations. American Journal of Botany. 87(12): 1797-1806. 87:1797–1806.

Seifert, G. J., and C. Blaukopf. 2010. Irritable Walls: The Plant Extracellular Matrix and Signaling. Plant Physiology 153:467–478.
Seiler, J. R., and J. D. Johnson. 1988. Physiological and Morphological Responses of Three Half-Sib Families of Loblolly Pine to Water-Stress Conditioning. Forest Science 34:487–495.
Shi, H., L. Chen, T. Ye, X. Liu, K. Ding, and Z. Chan. 2014. Modulation of auxin content in Arabidopsis confers improved drought stress resistance. Plant Physiology and Biochemistry 82:209–217.
Taeger, S., T. H. Sparks, and A. Menzel. 2015. Effects of temperature and drought manipulations on seedlings of Scots pine provenances. Plant Biology 17:361–372.
UniProt: a hub for protein information. 2015. . Nucleic Acids Research 43:D204–D212.
Voltas, J., M. R. Chambel, M. A. Prada, and J. P. Ferrio. 2008. Climate-related variability in carbon and oxygen stable isotopes among populations of Aleppo pine grown in common-garden tests. Trees 22:759–769.
Wahid, N., S. C. González-Martínez, I. El Hadrami, and A. Boulli. 2006. Variation of morphological traits in natural populations of maritime pine (Pinus pinaster Ait.) in Morocco. Annals of forest science 63:83–92.
White, T. L., W. T. Adams, and D. B. Neale. 2007. Forest Genetics. CABI.
Williams, C. G., editor. 2009. The Dynamic Wind-Pollinated Mating System. Pages 125–135 Conifer Reproductive Biology. Springer Netherlands, Dordrecht.
Wolf, S., and S. Greiner. 2012. Growth control by cell wall pectins. Protoplasma 249:169–175.
Xing, H., X. Fu, C. Yang, X. Tang, L. Guo, C. Li, C. Xu, and K. Luo. 2018. Genome-wide investigation of pentatricopeptide repeat gene family in poplar and their expression analysis in response to biotic and abiotic stresses. Scientific Reports 8:1–9.
Xu, W., L. Jia, W. Shi, J. Liang, F. Zhou, Q. Li, and J. Zhang. 2013. Abscisic acid accumulation modulates auxin transport in the root tip to enhance proton secretion for maintaining root growth under moderate water stress. New Phytologist 197:139–150.

Table 4.1 Definition of 9 phenotypic traits and the ANOVA analysis results for them between wet and dry treatment, including RL, SW, RW, SRL, R2S, SDAD, NRAD, NRAB, and GR.

Trait	Definition	*p*_value
RL	Root length (cm)	4.58e-08 ***
SW	Shoot weight (g)	2.18e-08 ***
RW	Root weight (g)	0.48846
SRL	Specific root length (cm/g)	0.0514
R2S	Root shoot dry mass ratio	1.12e-11 ***
SDAD	Stomata density on adaxial side	2.00e-14 ***
NRAD	Number of stomata row on adaxial side	0.1841
NRAB	Number of stomata row on abaxial side	0.0225 *
GR	Height growth (cm)	2e-16 ***

* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$.

Table 4.2 SNP annotation with SnpEFF for SNPs significantly associated with Root Length (RL), Shoot Weight (SW), Stomata density on adaxial side (SDAD), Number of stomata row on abaxial side (NRAB), Root shoot dry mass (R2S), and Height growth (GR).

Variant type	RL	SW	SDAD	NRAB	R2S	GR
intergenic	104	67	27	70	1317	113
downstream	3	4	1	1	21	7
intragenic	35	9	5	9	131	12
synonymous	3	0	2	1	12	1
upstream	5	0	4	4	35	8
missense	3	0	3	0	14	4
Total	153	80	42	85	1530	145

Table 4.3 Gene ontology for selected phenotypic-associated SNPs

SNP	Variant type (gene ID)	Function	Associated phenotypic variables
V162712	intragenic (PHYPA_023122)	Ubiquitination	RL
V3237504	upstream (C4D60_Mb07t09060)	Ubiquitination	RL
V3229603	upstream (LOC104592768)	Ubiquitination	R2S
V75426	intragenic (UPL1)	Ubiquitination	GR
V604584	intragenic (UBC23)	Ubiquitination	R2S
V1267599	intragenic (RPN7)	Deubiquitination	SDAD
V3880001	intragenic (RPN2A)	Deubiquitination	SDAD
V3678861	missense (CEY00_Acc20295)	embryo development ending in seed dormancy	GR
V381864	downstream (ALDH7B4)	ABA response; desiccation response	NRAB
V2990435	missense (WRKY51)	ABA response	R2S
V979582	upstream (CRK2)	ABA signaling	R2S
V114623	intragenic (DRP2B)	ABA reponse; root hair initiation	R2S
V2262698	downstream (RchiOBHm_Chr6g0256501)	transferase activity (UDP-glucosyltransferase protein involved in water stress response through IBA)	R2S
V2183716	synonymous (ERF094)	ethylene-activated signaling pathway; response to ethylene and jasmonic acid	RL
V765589	downstream (CSN5A)	COP9 signalosome assembly (COP9 involved in auxin and jasmonate responses)	R2S

V1290603	synonymous (MIMGU_mgv1a012336mg)	Cell wall organization	NRAB
V1412402	missense (TBR)	Cell wall modification	R2S
V193982	intragenic (PME53)	Cell wall modification; pectin catabolic process	R2S
V268065	intragenic (PAE6)	Cell wall modification; pectin acetylesterase activity	R2S
V1766242	intragenic (ABCB4)	transport of auxin in roots; root hair elongation	R2S
V1335792	intragenic (YUC9)	auxin biosynthesis in roots	SW
V3570404	missense (TRN1)	auxin-activated signaling pathway; leaf development	R2S
V4107103	upstream (LBD18)	lateral root formation; xylem development	NRAB
V328590	downstream (MKK6)	lateral root formation; stress-activated protein kinase signaling cascade; signal transduction by protein phosphorylation	R2S
V148025	upstream (KEU)	vesicle trafficking (involved in root hair development); regulation of defense response	R2S
V1906092	missense (At5g63930)	protein serine/threonine kinase activity (leucine-rich protein involved in stress in roots)	R2S
V701195	synonymous (At1g33600)	hormone-mediated signaling pathway (leucine-rich protein involved in stress in roots)	R2S
V113920	upstream (BVC80_9003g22)	hydrolase activity (protein Inosine-uridine preferring nucleoside hydrolase response to drought in roots)	R2S
V296113	upstream (ACMD2_00993)	hydrolase activity (expression of Metallophos protein during drought)	R2S

V1479332	intragenic (NA)	zinc ion binding (Retrovirus-related Pol polyprotein from transposon TNT 1-9 related to biotic and abiotic stress)	R2S
V1601722	intragenic (PCMP-E76)	zinc ion binding (Pentatricopeptide repeat-containing protein linked to abiotic and biotic stress)	R2S
V2674615	missense (AMTR_s00060p00214220)	signal transduction	SDAD
V3159668	missense (TIR)	defense response; signal transduction	R2S
V1862744	downstream (CPK34)	pollen tube growth	R2S
V1647269	intragenic (PCMP-E76)	lignin biosynthetic process; zinc ion binding	R2S
V4145513	upstream (CYCA2-2)	cell division; stomatal complex development	GR

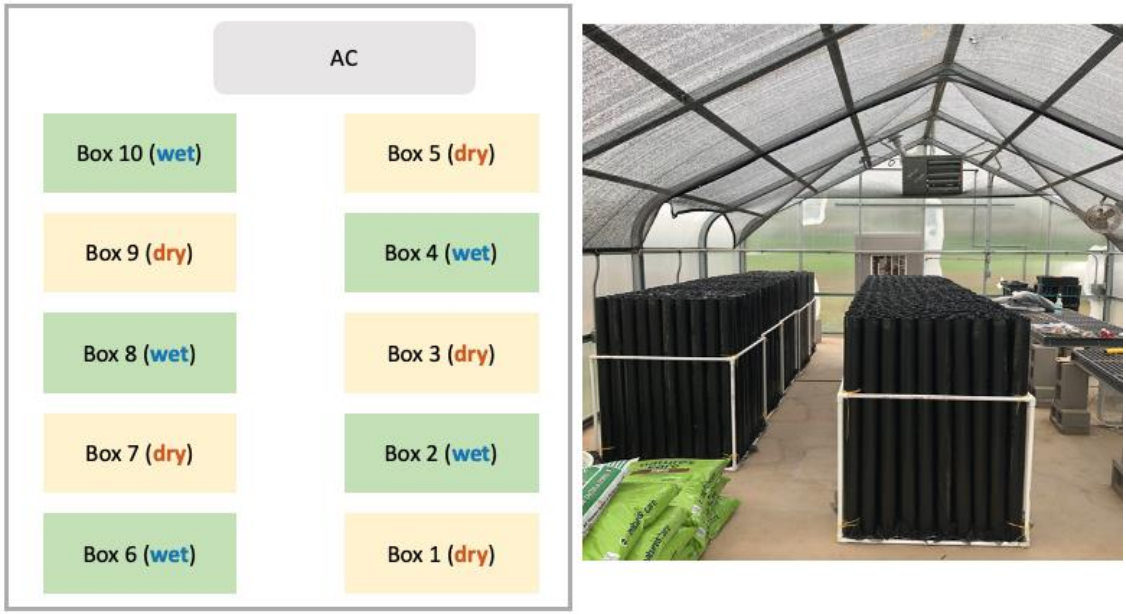

Figure 4.1 Greenhouse setup for 10 boxes of tubes, including five wet and five drought ones. There are 100 seedling tubes in each of the box.

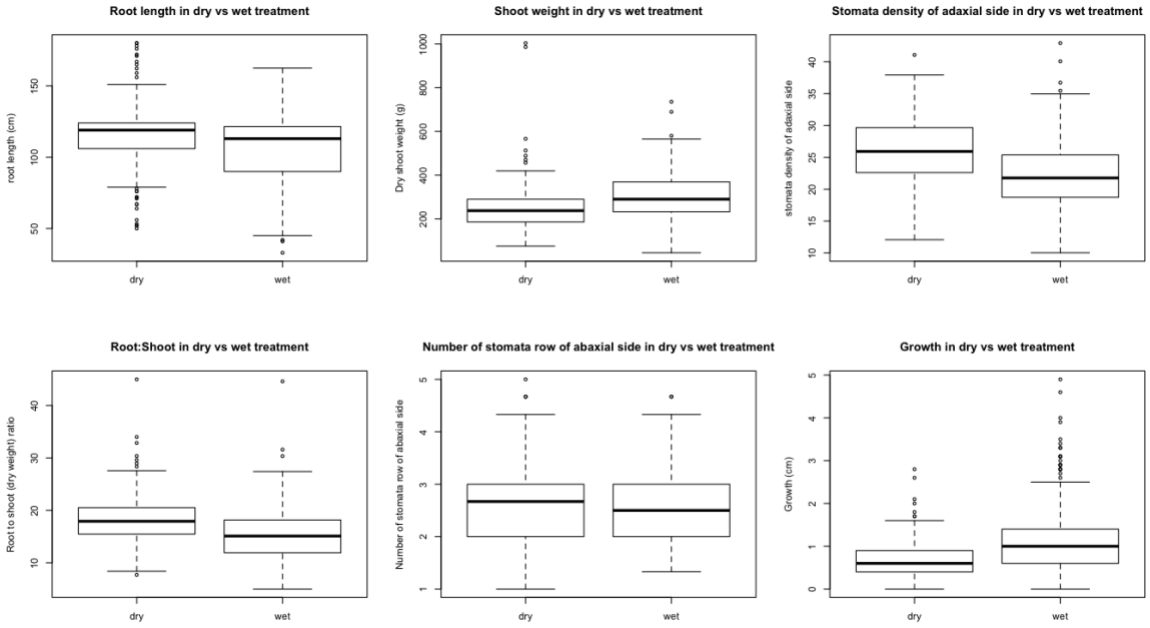

Figure 4.2 Boxplot of 6 traits in wet and dry treatment, including Root Length (RL), Shoot Weight (SW), Stomata density on adaxial side (SDAD), Root shoot dry mass (R2S), Number of stomata row on abaxial side (NRAB), and Height growth (GR).