EXOME GENOTYPING AND ASSOCIATION GENETICS OF QUANTITATIVE TRAITS IN A CLONALLY TESTED LOBLOLLY PINE POPULATION (*PINUS TAEDA* L.)

A Dissertation

by

MENGMENG LU

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

DOCTOR OF PHILOSOPHY

Chair of Committee,	Carol A. Loopstra
Co-Chair of Committee,	Konstantin V. Krutovsky
Committee Members,	Thomas D. Byram
	Alan E. Pepper
	Shuhua Yuan
Intercollegiate Faculty Chair,	Dirk B. Hays

December 2016

Major Subject: Molecular and Environmental Plant Sciences

ABSTRACT

Loblolly pine (*Pinus taeda* L.) is one of the most widely planted and commercially important forest tree species in the USA and worldwide. However, whole genome resequencing in loblolly pine is hampered by its size and complexity. Additionally, the genetics underlying quantitative traits of loblolly pine remains to be discovered. As a valid and more feasible alternative, entire exome sequencing was hence employed to identify the gene-associated single nucleotide polymorphisms (SNPs) and to genotype the 375 tress in a clonally tested loblolly pine population. Adaptive and growth traits were also measured and analyzed on this population.

The exome capture efficiency was high. A total of 972,720 high quality SNPs were identified after filtering. We found that linkage disequilibrium (LD) decayed very rapidly within this population. Two main distinct clusters representing western and eastern parts of the loblolly pine range were demonstrated by the population structure analysis using unlinked SNPs. Under a relaxed filtering condition, over 2.8 million SNP markers were used to test for single locus associations, SNP-SNP interactions and correlation of individual heterozygosity with phenotypic traits. Genetic correlations between traits as well as geographical variation exist within this population. A total of 36 SNP-trait associations were found for specific leaf area (5 SNPs), branch angle (2), crown width (3), stem diameter (4), total height (9), carbon isotope discrimination (4), nitrogen concentration (2), and pitch canker resistance (7). Eleven SNP-SNP interactions were found to be associated with branch angle (1 SNP-SNP interaction), crown width

(2), total height (2), carbon isotope discrimination (2), nitrogen concentration (1), and pitch canker resistance (3). Non-additive effects imposed by dominance and epistasis compose a large fraction of the genetic variance for the quantitative traits. Candidate genes that underlie these traits have a wide spectrum of functions.

The obtained results demonstrated the efficiency of exome capture for genotyping species such as loblolly pine with a large and complex genome. Multiple effects that influence the performance of loblolly pines identified in this study provide great resources for understanding the genetic control of complex traits, and have potential value for breeding through maker assisted selection and genomic selection.

DEDICATION

This dissertation is dedicated to my dear parents!

ACKNOWLEDGEMENTS

I would like to express the deepest gratitude to my graduate advisors Dr. Carol Loopstra and Dr. Konstantin (Kostya) Krutovsky for their constant encouragement and excellent guidance through the course of my doctoral program. I thank their valuable help and unlimited patience regarding my research and dissertation writing. I am also grateful to the other members of my graduate committee, Dr. Thomas Byram, Dr. Alan Pepper and Dr. Shuhua (Joshua) Yuan for their support of my graduate work.

Special thanks go to Jeff Puryear for his assistance and support during my Ph.D. years. I thank Dr. Claudio Casola for discussion and suggestions on the analysis methods. I would like to thank my friends, colleagues and the faculty and staff of the Program of Molecular and Environmental Plant Sciences (MEPS) and the Department of Ecosystem Science and Management (ESSM) for making my time at Texas A&M University a great experience.

This dissertation would not have been accomplished without the generous funding by the Pine Integrated Network: Education, Mitigation, and Adaptation Project (PINEMAP), a Coordinated Agricultural Project funded by the USDA National Institute of Food and Agriculture, Award #2011-68002-30185. Thanks to MEPS Travel Award funding my conference trips.

My thanks go to the PINEMAP team members Dr. C. Dana Nelson, Dr. Tomasz Koralewski, Dr. Jason Vogel, Dr. Jason West, Dr. Ross Whetten for their mentorship and contribution to this dissertation. I am deeply grateful to Dr. Vikram Chhatre, Dr.

Fred Raley, Nathalie Reilly and technicians (Gay Flurry, Chance Parker, Chuck Burdine) of the Harrison Experimental Forest at Southern Institute of Forest Genetics for their help on sample collection.

I also want to extend my thanks to Dr. Jill Wegrzyn, John Liechty, Hans Vasquez-Gross and PineRefSeq Project team for providing the draft loblolly pine reference sequences, exon annotation and bioinformatics assistance, Dr. John Davis and Dr. Tania Quesada for providing pitch canker resistance data, Texas A&M Institute for Genome Sciences and Society (TIGSS) for providing computational resources and system administration support for the TIGSS HPC Cluster.

Finally, I could not make this far without the love, encouragement and support from my mother, father, other members in the family and many friends. You made me who I am!

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Carol Loopstra (Department of Ecosystem Science and Management), Dr. Konstantin Krutovsky (adjunct faculty at Texas A&M University, full professor of the Department of Forest Genetics and Forest Tree Breeding at Georg-August-University of Göttingen), Dr. Thomas Byram (Department of Ecosystem Science and Management), Dr. Alan Pepper (Department of Biology) and Dr. Shuhua (Joshua) Yuan (Department of Plant Pathology and Microbiology).

The loblolly pine draft reference sequences and annotation data were provided by Dr. Jill Wegrzyn (Department of Ecology and Evolutionary Biology, University of Connecticut) and PineRefSeq Project team. The pitch canker resistance data analyzed for the third section were provided by Dr. John Davis and Dr. Tania Quesada (School of Forest Resources and Conservation, University of Florida). The analyses depicted in the second section were conducted by the student in collaboration with Dr. Konstantin Krutovsky, Dr. Tomasz Koralewski (Department of Ecosystem Science and Management), Dr. Thomas Byram, Dr. C. Dana Nelson (USDA Forest Service, Southern Research Station, Southern Institute of Forest Genetics) and Dr. Carol Loopstra. This part has been published in 2016 in an article listed in the *BMC Genomics*. The analyses depicted in the third section were conducted by the student in collaboration with Dr. Konstantin Krutovsky, Dr. C. Dana Nelson, Dr. Jason West (Department of Ecosystem

Science and Management), Nathalie Reilly (Department of Biology and Marine Biology, University of North Carolina Wilmington) and Dr. Carol Loopstra.

Funding Sources

Research and graduate study were supported by the Pine Integrated Network: Education, Mitigation, and Adaptation Project (PINEMAP), a Coordinated Agricultural Project funded by the USDA National Institute of Food and Agriculture, Award #2011-68002-30185.

NOMENCLATURE

2014H	Total height in the year of 2014
2015HB	Total height in the year of 2015 before the growing season
2015HA	Total height in the year of 2015 after the growing season
ABA	Phytohormone abscisic acid
ANOVA	Analysis of variance
ADEPT2	Allele Discovery of Economic Pine Traits 2
BA	Branch angle
BLUP	Best linear unbiased prediction
CDS	Coding sequences
$\Delta^{13}C$	Carbon isotope discrimination
CW	Crown width
DIA	Stem diameter
EST	Expressed sequence tag
GBS	Genotyping-by-sequencing
GLM	General linear model
HWE	Hardy-Weinberg equilibrium
HTC	Heterozygosity-trait correlation
LD	Linkage disequilibrium
MAF	Minor allele frequency
MLM	Mixed linear model

NCBI	National Center for Biotechnology Information
NIFA	National Institute of Food and Agriculture
NIT	Nitrogen concentration
NSF	National Science Foundation
PINEMAP	Pine Integrated Network: Education, Mitigation, and Adaptation Project
RAD-Seq	Restriction site associated DNA sequencing
SLA	Specific leaf area
sHSP	Small heat shock protein
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
T_S	Transition
T_V	Transversion
Ts/Tv	Transition to Transversion ratio
UTRs	Untranslated regions
VPD	Vapour pressure deficit
WUE	Water use efficiency

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Page

1. INTRODUCTION

Southern forests dominated by pines (genus *Pinus*) contain one third of the entire forest carbon in the contiguous U.S. (Turner et al. 1995). Among the southern pines, loblolly pine (*P. taeda* L.) is the most common native forest tree species, providing great economical and ecological value to this area. It comprises nearly one-fourth or 55 million acres of the southern forest in the U.S. (Smith et al. 2009).The native range of loblolly pine extends south from New Jersey to central Florida, and west to central Texas (Baker and Langdon 1990). Its amenability to plantation management, high yields and fast growth make it one of the most economically important forest species in the world. Timber and pulpwood are the primary products. Due to its rapid juvenile growth, loblolly pine is the most productive and valuable commercial forest species, comprising 80% of the planted forestland and over one half of the standing volume in the southern U.S. (Huggett et al. 2013). Since forests capture and store carbon dioxide through photosynthesis, the widely planted loblolly pines provide great value in offsetting atmospheric carbon dioxide and mitigating long-lasting climate changes caused by greenhouse gas emissions (Millar et al. 2007; Bolte et al. 2009).

The climate within the natural loblolly pine range is humid, warm-temperate with long, hot summers and mild winters (Baker and Langdon 1990). However, due to climate change, it may become more arid and threaten survival and growth, hence decreasing the productivity and the economic value of loblolly pine cultivation. Forest growth, along with crop yield, can be damaged by changing precipitation and extreme weather, such as heat waves, drought, floods, and high winds (IPCC 2014b). For instance, during the severe droughts in 2005 and 2010, the Amazon lost a decade's worth of carbon sequestration (Bellassen and Luyssaert 2014). Heat waves due to climate change were attributed as the main reason for longer fire seasons in Russia and Alaska. In 2010, 2.3 million hectares of Russian forests were affected by a record heat wave with forest fires (Bellassen and Luyssaert 2014). In 2015, over 2 million hectares of forests burned in Alaska. Warmer temperatures in Alaska are accelerating the melting of permafrost, and tons of carbon may be released (Zhang 2015). On the other hand, forests are the primary mechanism for carbon sequestration (Bellassen and Luyssaert 2014). Thirty percent of annual global anthropogenic CO_2 emissions have been absorbed by the world's forest in the past few decades. Moreover, atmospheric records and forest inventories show that forests have been taking up more CO_2 in the past 50 years with the growing trend of anthropogenic emissions (Bellassen and Luyssaert 2014). Therefore, adaptive forest management that takes into account the changing environment due to climate change could relieve warming and long-lasting climate changes (Millar et al. 2007; Bolte et al. 2009).

In the southeastern U.S., climate change poses challenges for pine adaptation. According to a 2009 summary from the U.S. Global Change Research program (Global climate change impacts in the United States 2009), continued temperature increases, high heat in summer and precipitation declines are predicted in the southeastern U.S. It is assumed that pine carbon sequestration will be negatively impacted by soil water deficits under these conditions (Noormets et al. 2010). Over the past few decades, research cooperatives working to improve pine management and genetics have tripled the productivity of planted pine (Jokela et al. 2010). In the face of changing climate, the development and deployment of improved genetics, seedling culture, and nutrient management technology will play important roles in pine adaptation, resilience and sustainability.

In 2011, a grant funding the Pine Integrated Network: Education, Mitigation, and Adaptation Project (PINEMAP) was awarded to a large multi-institutional and multidisciplinary group of scientists by the USDA National Institute of Food and Agriculture (NIFA). The overarching goal of PINEMAP is to integrate research, extension, and education and to disseminate knowledge about forestry management, fertilizer efficiency and carbon sequestration to southern pine landowners. PINEMAP has six aims (http://www.pinemap.org/) including a genetics team that is developing guidelines to help growers understand where to plant specific southern pine seed sources given future climate scenarios and identify genes controlling traits such as growth, nitrogen responsiveness, cold hardiness, water usage, and resistance to southern pine beetle and fungal diseases. As part of the genetics group, we have analyzed the genetics and genomics of a clonally tested association mapping population to discover alleles in genes that control these important adaptation and productivity traits. The population consisting of rooted cuttings from unrelated trees (i.e., clones) sampled across the natural range of loblolly pine was previously developed and used for association analyses as part of the National Science Foundation (NSF) funded Allele Discovery of Economic Pine Traits 2 (ADEPT2) project (Eckert et al. 2010; Cumbie et al. 2011). In the spring of

2010, 384 of the clones were established at the Harrison Experimental Forest at the Southern Institute of Forest Genetics (Saucier, Mississippi) and those trees were utilized for this project.

The goal of this research was to discover single nucleotide polymorphisms (SNPs) in the 384 trees in the ADEPT2 population. The genotyped SNPs were used to analyze the population genetics and identify the genes and alleles that control adaptive and productivity traits.

Endeavors have been made to improve discovery of genetic markers related to genes of interest for loblolly pine in the past. In the ADEPT2 project, 7216 SNPs were selected to design an Illumina Infinium middle density SNP genotyping assay for association analyses (Eckert et al. 2009a). Approximately 4000 of them provided high quality and reliable genotypes that were used to associate SNPs with variation for pitch canker resistance, carbon isotope discrimination, nitrogen concentration, height, metabolite levels and gene expression (Quesada et al. 2010; Cumbie et al. 2011; Eckert et al. 2012; Palle et al. 2013). However, this relatively low number of SNPs was insufficient to fully dissect the genetic structure of the complex traits. In addition, the complexity of the loblolly pine genome and the lack of a complete reference genome created challenges for marker discovery at that time. With the advances in sequencing technology, genotyping-by-sequencing (GBS) is now possible in species with high diversity and large genomes (Elshire et al. 2011). The advantage of GBS is that genetic variation discovery and genotyping can be completed simultaneously, hence saving time and money. In June 2012, the first draft reference assembly of loblolly pine was released (Neale et al. 2014), which facilitated marker discovery via GBS.

The loblolly pine genome with a size of ≈ 23 Gbp is still a challenge for whole genome resequencing. Therefore, reduction of the genome complexity is required for application of GBS in loblolly pine. There are two major GBS library construction methods: target enrichment and restriction enzyme based methods. In the PINEMAP project, a team from North Carolina State University constructed GBS libraries based on reducing genome complexity using RAD-Seq method (restriction site associated DNA sequencing), while our team at Texas A&M University and one from the University of Florida utilized target enrichment methods.

One advantage of target enrichment lies in the reduction of the amount of sequencing required and the difficulties in data analyses. Since the loblolly pine genome is characterized by divergent and abundant repetitive elements (Kovach et al. 2010; Wegrzyn et al. 2014), targeting and genotyping the exome instead of the whole genome could save money and time but still uncover a significant amount of variation in most coding sequences (CDS) and untranslated regions (UTRs). Different exome capture systems have been developed in the past few years. Among them, the Agilent SureSelect Target Enrichment (Agilent Technologies, Inc., Santa Clara, CA) and NimbleGen SeqCap EZ Systems (Roche NimbleGen, Inc., Madison, WI) support solution-based capture of regions of customer interest. As shown in Figures 1.1 and 1.2, both methods share a similar procedure. First, genomic DNA is sheared to small fragments, followed by end repair, A-tailing, adapter ligation and amplification steps. Second, the sample libraries are hybridized with the designed oligonucleotide probes (baits). Third, the captured sequences are pulled down and amplified after wash and recovery steps.

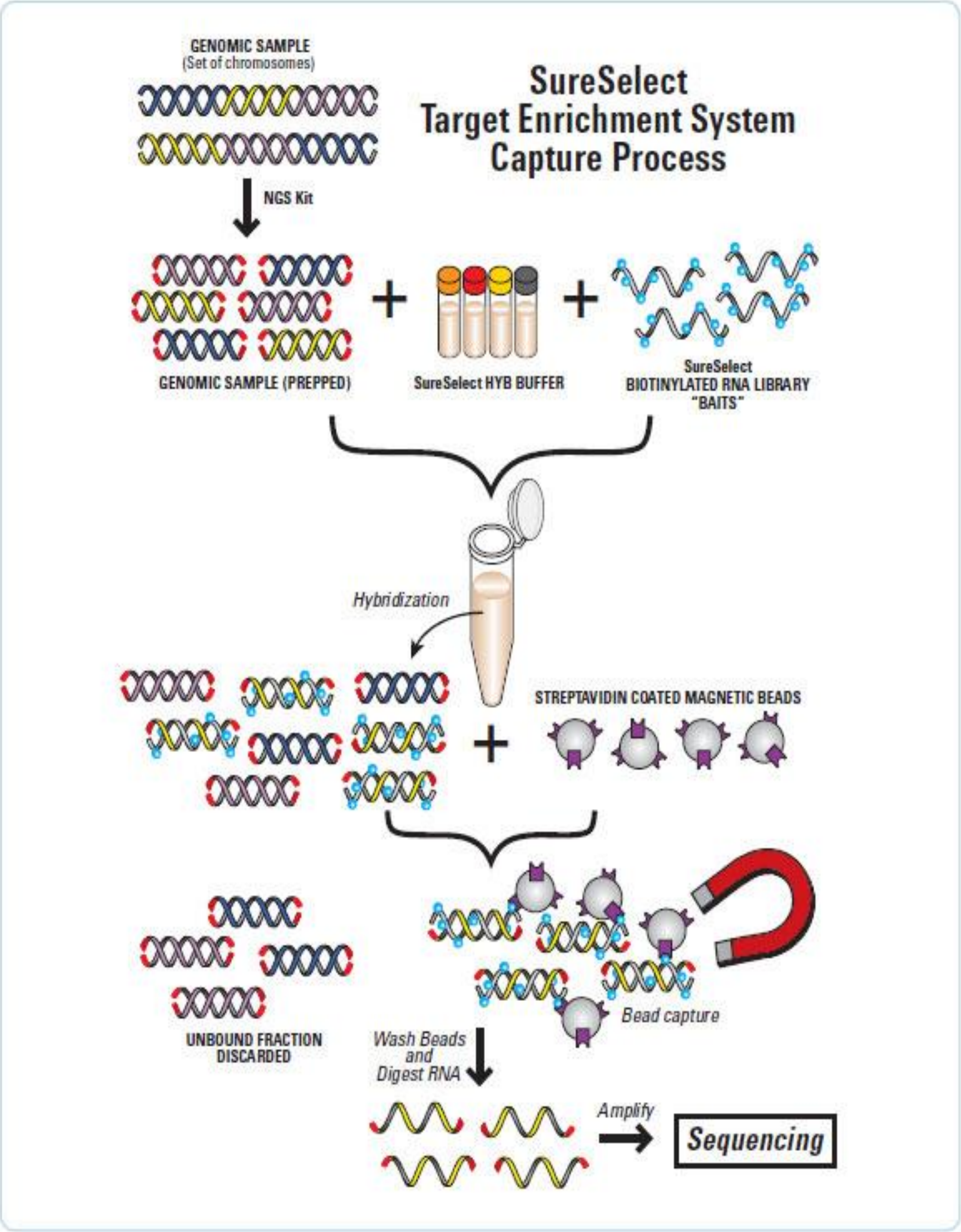

Fig. 1.1 Workflow of the Agilent SureSelect Target Enrichment System (http://www.genomics.agilent.com/article.jsp?pageId=3083)

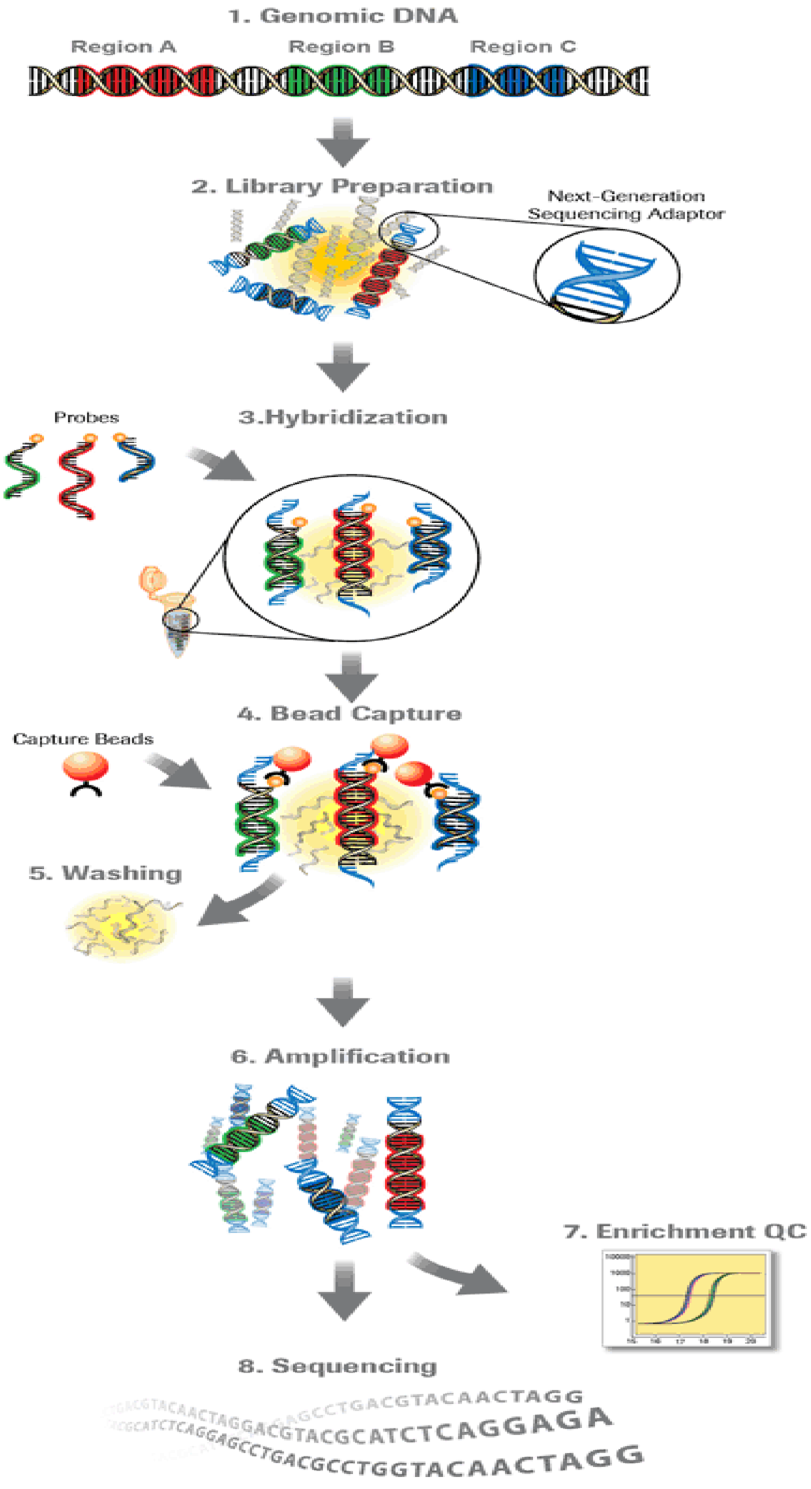

Fig. 1.2 Workflow of the NimbleGen SeqCap EZ System (http://www.nimblegen.com/products/seqcap/ez/choice/index.html)

There are some dissimilarities between these two methods. First, the oligonucleotide probes supplied by NimbleGen are DNA single strands with a flexible length of 55-105 bp while Agilent probes are strictly 120 bp long RNA strands. Second, the hybridization time for the Agilent system is 24 hours, while the NimbleGen system requires 68-72 hours.

In our pilot experiment, both the Agilent SureSelect Target Enrichment and NimbleGen SeqCap EZ Systems were tested to select and capture the loblolly pine exome. For the Agilent SureSelect Target Enrichment method, the baits were designed using the 35,550 unigenes (available on http://bioinfolab.muohio.edu/txid3352v1) assembled by Dr. Chun Liang (Miami University, Oxford, Ohio). These unigenes were assembled using the loblolly pine Expressed Sequence Tags (ESTs) deposited in NCBI dbEST. For the NimbleGen SeqCap EZ method, we designed the probes using 199,723 exons identified in the loblolly pine reference genome (http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/) (Wegrzyn et al. 2014). The capture efficiency and cost were compared to determine the best system to be applied for genotyping of this study.

From the pilot experiment, we found the capture efficiency with the NimbleGen system was higher than with the Agilent system. In addition, the cost was much lower with the NimbleGen SeqCap EZ system. Therefore, the NimbleGen SeqCap EZ system was selected to genotype the entire population. Afterwards, the genotyped SNPs were associated with adaptive and growth traits to discover molecular markers that can be used to facilitate future tree breeding.

In the second section, genotyping using NimbleGen SeqCap EZ system, SNP discovery and population genetics analyses on a loblolly pine population will be presented and discussed. In the third section, genetic correlations between phenotypic traits, geographic variation, SNPs and SNP-SNP interactions associated with the adaptive and growth traits will be presented and discussed.

2. EXOME GENOTYPING, LINKAGE DISEQUILIBRIUM AND POPULATION STRUCTURE IN LOBLOLLY PINE (*PINUS TAEDA* L.)[*]

2.1 Introduction

Genomic tools and resources that focus on the dissection of complex traits are revolutionizing traditional loblolly pine breeding and assist with the breeding and deployment of genotypes better adapted to climate change and able to sequester greater amount of carbon. Two key prerequisites for development and application of genomics-assisted breeding are the characterization of the genetic variation and the collection of genome-wide molecular markers. A high level of genetic polymorphism is expected in loblolly pine due to its life traits, typical for conifer species, such as longevity, wide geographic distribution, large effective population size and high outcrossing rate. This was confirmed in early studies with isozymes (Adams and Joly 1980; Conkle 1981), DNA-based markers (Devey et al. 1994; Harry et al. 1998; Remington et al. 1999), especially more recently with SNP (Brown et al. 2004; Eckert et al. 2009b; Chhatre et al. 2013) markers. About 4000 SNP markers have been genotyped in the previous association genetics studies (Eckert et al. 2009b; Cumbie et al. 2011), but many more markers are needed for genomic selection (Jannink et al. 2010; Resende Jr et al. 2012; Resende et al. 2012; Desta and Ortiz 2014).

[*] Reprinted with permission from "Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.)" by Lu M, Krutovsky KV, Nelson CD, Koralewski TE, Byram TD and Loopstra CA, 2016. *BMC Genomics*, DOI: 10.1186/s12864-016-3081-8, Copyright [2016] by Lu et al.

In the previous loblolly pine association mapping studies, an Illumina Infinium high-throughput SNP genotyping array developed for multiplex genotyping of 7216 SNP markers was used to dissect genetic control of diverse phenotypic traits (Eckert et al. 2009b; Eckert et al. 2010; Quesada et al. 2010; Cumbie et al. 2011; Chhatre et al. 2013; Palle et al. 2013). These SNPs were derived originally from amplicon sequencing data based on a relatively small, but range-wide sample of 18 loblolly pine megagametophytes and using PCR primers that were designed using unigene contig sequences assembled from EST sequences. Finally, about 4000 SNPs from this 7K SNP array were polymorphic or could be genotyped in the follow-up studies (Eckert et al. 2009b; Eckert et al. 2010; Quesada et al. 2010; Cumbie et al. 2011; Chhatre et al. 2013; Palle et al. 2013).

Given adequate geographic distribution sampling, the genetic structure underlying loblolly pine populations could also be elucidated using SNPs. For instance, Eckert et al. (Eckert et al. 2010) analyzed SNP and simple sequence repeat (SSR) markers among 907 range-wide loblolly pine trees and found that the population structure reflected mainly the Mississippi River discontinuity.

Efficiency of marker-assisted breeding and genomic selection depends largely on genome-wide linkage disequilibrium (LD). Brown et al. (Brown et al. 2004) found substantial historic recombination between SNPs in the sampled alleles sequenced in 19 genes and demonstrated that LD significantly declined within 2 Kbp in loblolly pine. A genome-wide study by Chhatre et al. (Chhatre et al. 2013) confirmed rapid LD decay in loblolly pine. These studies suggested that a very large number of markers would be required to link phenotypes to genotypes in association mapping studies and in genomic selection of this species. Therefore, for a species such as loblolly pine with a large genome and rapid LD decay, even thousands of markers cannot meet the requirement of identifying all important functional genomic regions. Fortunately, GBS, which enables simultaneous marker discovery and genotyping, has facilitated the generation of large numbers of molecular markers (Poland and Rife 2012). Nevertheless, the large size and complex structure of the loblolly pine genome pose challenges for the whole genome resequencing. The loblolly pine genome assembly v. 1.01 spans 23.2 Gbp and contains 14.4 million scaffolds (Neale et al. 2014). Tentatively, 50,172 putative genes with an average length of 2.7 Kbp have been annotated in the current loblolly pine genome assembly (Wegrzyn et al. 2014). Moreover, various highly repetitive DNA elements compose up to 82% of the loblolly pine genome, among which retrotransposons dominate and comprise 62 % of the genome (Neale et al. 2014; Wegrzyn et al. 2014). Therefore, reduction of genome complexity is highly desired for application of GBS to loblolly pine.

In our study, we used the entire exome region for target enrichment to limit GBS to mostly CDS, which represent only ~40-60 Mbp of sequence space or less than 0.2 % of the entire loblolly pine genome. In the previous studies, technologies for solution-based enrichment of target regions of interest have been developed for loblolly pine (Neves et al. 2011; Neves et al. 2013b; Neves et al. 2013a). Capture size has been significantly expanded due to the improvement in probe design and capture efficiency, making it possible to capture up to 200 Mbp of target sequence with a single design

(NimbleGen SeqCap EZ Developer Enrichment Kit). These developments made it possible for us to target and enrich the entire loblolly pine exome, thus greatly enlarging the available number of molecular polymorphisms in loblolly pine.

In this study, we describe the probe design and efficiency of the loblolly pine exome capture using the NimbleGen SeqCap EZ method in a population sample containing 375 clonally-propagated trees from an association mapping population generated for the ADEPT 2 project (Cumbie et al. 2011). Counties of origin are known for 362 out of 375 maternal trees (Fig. 2.1). SNPs were identified by aligning the exome capture sequences to loblolly pine genome assembly v. 1.01 (Zimin et al. 2014). The inferred SNP genotypes were then applied to study LD decay and population structure.

2.2 Methods

2.2.1 Plant Material and Genomic DNA Extraction

The population studied here was from the ADEPT2 project (Cumbie et al. 2011). Maternal parents of the ADEPT2 population were originally sampled across 12 states in the southeastern U.S., extending from Virginia to Florida, and west to central Texas (Fig. 2.1). Seeds were collected from the maternal trees after open pollination. Trees were grown from open-pollinated seeds for one year and then were hedged and established for use in the ADEPT2 project. In the spring of 2010, rooted cuttings from 384 trees (i.e., clones) of the ADEPT2 population were established at the Harrison Experimental Forest at the Southern Institute of Forest Genetics, near Saucier, Mississippi. Needle samples were collected from 375 surviving clones for extraction of genomic DNA in June 2012 and stored at -20°C. Four needles from each sample were ground in liquid nitrogen to a fine powder. DNA was extracted using QIAGEN DNeasy Plant Mini Kits following the standard protocol except in the last step, where 1×TE buffer with low EDTA was used for elution. Genomic DNA samples with OD260/OD280 ratios between 1.7 and 2.0 without signs of degradation were used for downstream applications.

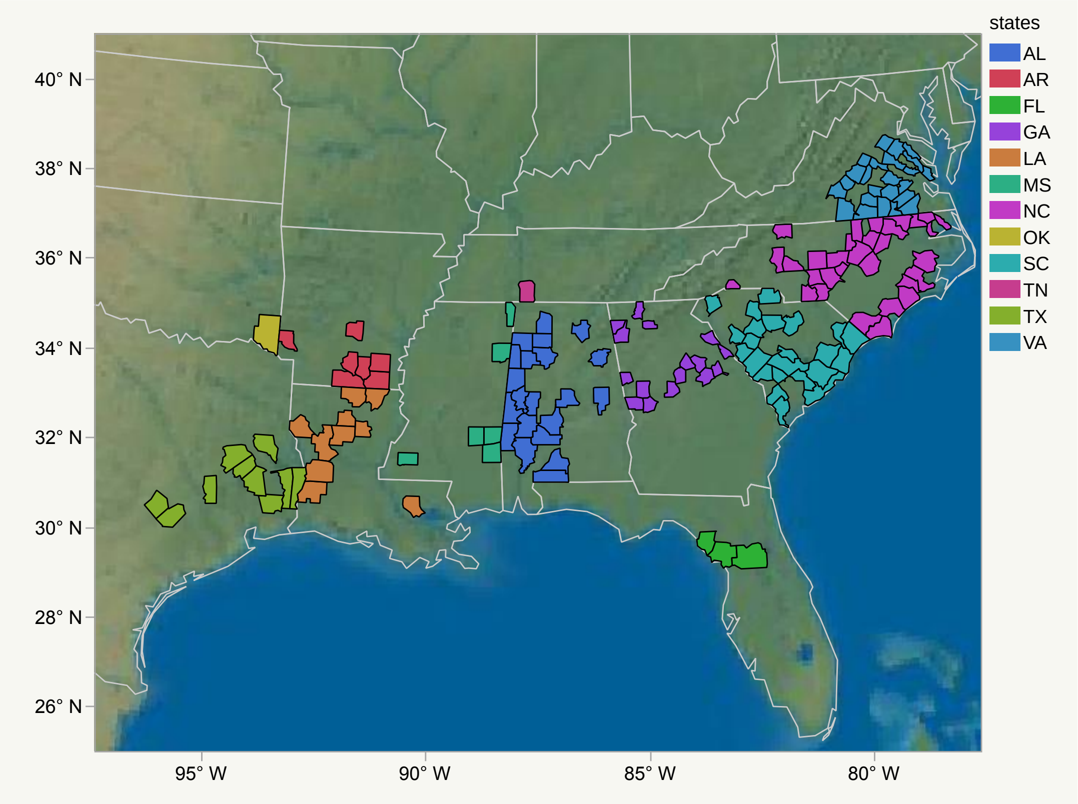

Fig. 2.1 The counties of origin of the maternal trees colored by states. This map shows the sampling sites of the 362 out of 375 maternal parents of the ADEPT2 population used in this study

2.2.2 Probe Design

Probes were designed using Gene Annotation v. 2.0 for loblolly pine genome assembly v. 1.01 (Neale et al. 2014; Wegrzyn et al. 2014). We submitted the 49,216,700 bp of sequence that represented 199,723 exons to Roche NimbleGen Inc. for sequence capture probe selection. The target regions were inferred using the exon coordinates available in the files "ptaeda.v1.01 scaffolds.trimmed.all.genes.highq_whole.gff3", which included annotation for 34,059 full length, high quality genes, and "ptaeda.v1.01 scaffolds.trimmed.all.genes.highq_partial.gff3", which included 14,332 partial length, high quality genes. Exons shorter than 100 bp in length were extended (padded) to 100 bp. After screening, a total of 196,068 exons (51,239,342 bp) were selected for probe design. A relatively conservative threshold was used to design unique probes that could tolerate no more than five single-base indel or single nucleotide substitution mismatches with the genome. The length of the probes varied between 50 and 100 bp. The average length was 76.5 ± 4.2 bp, with a median of 76 bp.

2.2.3 Sequencing Library Preparation and Target Enrichment

Each genomic DNA was diluted to 25 ng/µl in 1×TE buffer with low EDTA and 50 µl of each DNA solution was fragmented to have an average size distribution of ~180-220 bp using a Covaris sonicator. KAPA Library Preparation Kits (Illumina® Platforms) were used to construct a library for each DNA sample. After post-ligation cleanup and dual-SPRI size selection, the sample libraries were amplified and checked for quality and quantity using the Agilent 2100 Bioanalyzer and PicoGreen dsDNA quantitation assays. The amplified sample library was acceptable if the OD260/OD280 ratios were between 1.7 and 2.0, respectively, the yield was more than 1.0 μg, and the average fragment size was between 150 and 500 bp.

The Roche NimbleGen SeqCap EZ system was used for hybridization and target enrichment. Briefly, equal amounts of each of ten libraries representing uniquely individually indexed and amplified trees were mixed in a single exome enrichment and sequencing pool with a combined mass of at least 1.25 μg. The multiplexed paired-end sequencing libraries were hybridized with the target sequence capture probes and the mixture was incubated at 47°C for 72 hours. After wash and recovery steps, the captured multiplex DNA samples were amplified and purified. Following quality check, the captured multiplex DNA samples were loaded into Illumina HiSeq 2500 flowcells (one exome enriched pool of 10 original sample libraries per a single flowcell lane) and sequenced using 2×125 cycles at the Texas A&M University Genomics and Bioinformatics Service (College Station, Texas, USA).

2.2.4 Sequence Read Alignment and Analysis

Sequence reads for each of the 375 trees were filtered and demultiplexed. Then, the reads were mapped to loblolly pine genome assembly v. 1.01 (Neale et al. 2014; Wegrzyn et al. 2014; Zimin et al. 2014) using the “mem” routine in the BWA software v. 0.7.12 (Li and Durbin 2009) with the default parameters. The SAM files were converted to BAM files using the “view” routine in the SAMtools software v. 1.1 (Li et al. 2009). The “flagstat” routine in the SAMtools software was applied to calculate the mapping percentage of reads. The reads were filtered by the “view” and “sort” routines in the SAMtools software to acquire only the uniquely mapped and properly paired sorted reads. The “rmdup” routine in the SAMtools software was used to remove potential PCR duplicates from the filtered reads. The “intersect” routine in the BEDtools software v. 2.23.0-20-gada04b62.18 was applied to estimate the percentage of reads on target regions and the “coverage” routine was applied to visualize coverage of targeted DNA (Quinlan and Hall 2010).

Raw SNPs were called using the “mpileup” routine in the SAMtools software with 20 as the minimum mapping quality threshold for an alignment. The raw SNPs were filtered for downstream analyses, and only those that met the following criteria were kept: 1) 10× sequencing coverage in no less than 90% of all individuals. 2) bi-allelic; 3) minor allele frequency greater than 0.05. The VCFtools v. 0.1.12b software (Danecek et al. 2011) was applied to classify the SNPs according to their genomic regions and their positions relative to capture target regions. The SNP density was determined as the number of SNPs in a given region divided by the length of that regions.

2.2.5 Population Genetics Metrics

The VCFtools software was applied to calculate the minor allele frequency (MAF), the ratio of transition to transversion (Ts/Tv), individual heterozygosity and F_{IS}, and nucleotide diversity. The histogram graphs were plotted using the ggplot2 v. 2.1.0

package in R v. 3.2.3 (Wickham 2009; R Core Team 2015). The squared correlation coefficient between genotypes (r^2) on the same scaffold was used as an LD measure and calculated using the “geno-r2” routine in the VCFtools software. The trendline of LD decay along physical distance were fitted by nonlinear regression following Hill and Weir (Hill and Weir 1988). R software was applied to display the results (R Core Team 2015). The F_{ST} was estimated using the “weir-fst-pop” routine in the VCFtools software.

The SNP set was thinned to a single marker within every 1 Mbp distance in each scaffold” and converted to the PLINK software format using the “thin” and “plink” routines in the VCFtools software. The PLINK format was further converted to the PLINK BED format using the “make-bed” routine in the PLINK software v. 1.9 (Chang et al. 2015). The fastStructure software with the simple prior was applied to infer the most likely population structure by testing different number of potential subpopulations or clusters (K) from 2 to 12 (Raj et al. 2014). The recommended algorithm incorporated in fastStructure was applied to determine the reasonable choice of K. The admixture proportions of each individual were plotted using the Excel and R v. 3.2.3 (R Core Team 2015).

2.3 Results

2.3.1 Exome Target Enrichment Hybridization Probe Design and Assessment

Sequence capture oligonucleotide probes were designed using 199,723 exons in 48,391 (34,059 full-length and 14,332 partial-length) high quality tentative genes listed in gene annotation v. 2.0 for loblolly pine genome assembly v. 1.01 (Wegrzyn et al.

2014). The final probe set used in this study is available from Roche NimbleGen as custom SeqCap EZ design “140422_Ptaeda_Exome_ML_EZ_HX3”. Approximately 2.1 million single strand oligonucleotide probes were designed and produced in total that covered 90.2 % (46,206,684 bp) of the target regions. The regions not covered (gaps) were areas where the probe selection algorithm could not find a valid probe. These gaps usually represented repetitive DNA regions that, if included, could be expected to cause problems by capturing other homologous regions in the genome and, therefore, decrease capture and mapping efficiency.

2.3.2 Exome Capture Sequence Alignment and Efficiency

We multiplexed ten individually amplified and uniquely barcoded trees per library for capture hybridization, enrichment, and sequencing. After demultiplexing and filtering, we obtained between 25.25 and 60.55 million sequence reads per tree. The reads of each tree were mapped to loblolly pine genome assembly v. 1.01 (Neale et al. 2014; Wegrzyn et al. 2014; Zimin et al. 2014). Nearly 99 % of the sequence reads were mapped to the reference genome assembly. In order to improve the SNP discovery accuracy, the mapped reads were further filtered and only the uniquely mapped, properly paired (correctly oriented with respect to one another) and non-redundant reads were used for downstream analyses. After filtering, 62-75 % of the total reads (71 % per tree on average) were used for SNP calling.

Capture breadth and depth were investigated to examine capture efficiency and target specificity. For the uniquely mapped, properly paired, and non-redundant reads for each tree, we calculated the number of reads that mapped to the capture target regions using the BEDtools software v. 2.23.0-20-gada04b6 (Quinlan and Hall 2010). On average, 67 % of the reads per tree (59-74 %) mapped to the capture target regions. Additional non-target captured sequences included those adjacent to target or homologous regions. Between 91 % and 95 % of the capture target regions were covered by at least one read. The number of covered capture target bases was weakly and positively correlated with an increase in sequencing output (Fig. 2.2a; $r^2 = 0.23$, $P < 0.001$).

Coverage depth among the 375 trees was generally uniform and it was consistent across target regions. Among all the trees, at least 83 % of the capture target bases had coverage of 5X, 72 % - 10X, and 49 % - 20X (Fig. 2.3). The number of target bases with coverage depth of 10X or greater (Fig. 2.2b) seemed to change approximately linearly within a limited range of the total number of reads at about 37–55 million. Below this range, the number of captured bases increases faster than within the range. But the effect of increasing becomes weaker above 55 million. The mean coverage depth (Fig. 2.2c) increased linearly as the total sequencing output increased ($r^2 = 0.72$, $P < 0.001$), although the variance seemed slightly increased for the lower numbers of the total number of reads.

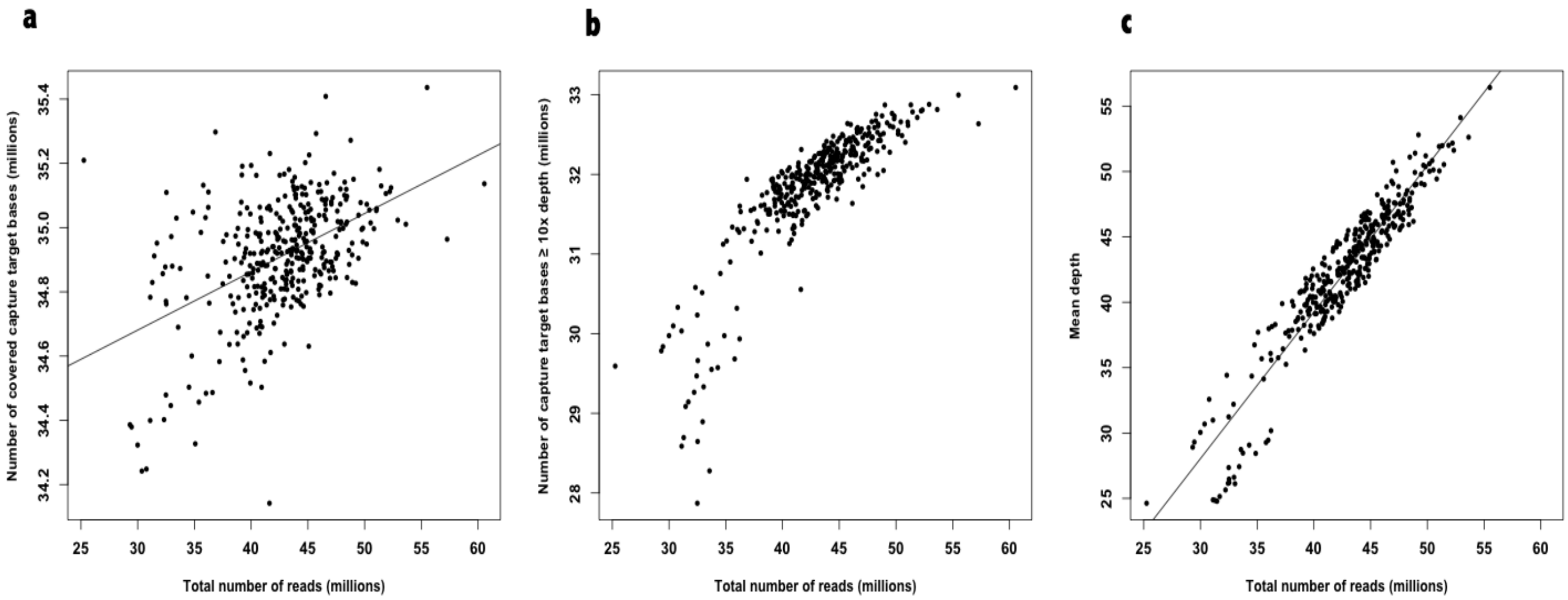

Fig. 2.2 Relationship between reads and capture target bases. a Relationship between reads and numbers of covered capture target bases. The numbers of captured target nucleotide bases are plotted against total number of sequence reads obtained in 375 trees from exome capture sequencing. The linear regression coefficient (r^2) is 0.23 ($P < 0.001$). b Distribution of on-target coverage ≥ 10X depth across the 375 trees. The numbers of capture target bases with a coverage depth of ten or greater sequence reads per target are plotted against the total number of sequence reads. The relationship seemed to change approximately linearly within a limited range of the total number of reads at 37–55 million. c Distribution of mean coverage depth across the 375 trees. The mean coverage depth is plotted against the total number of sequence reads. The linear regression coefficients (r^2) was significant ($P <$ 0.001) and equaled 0.72

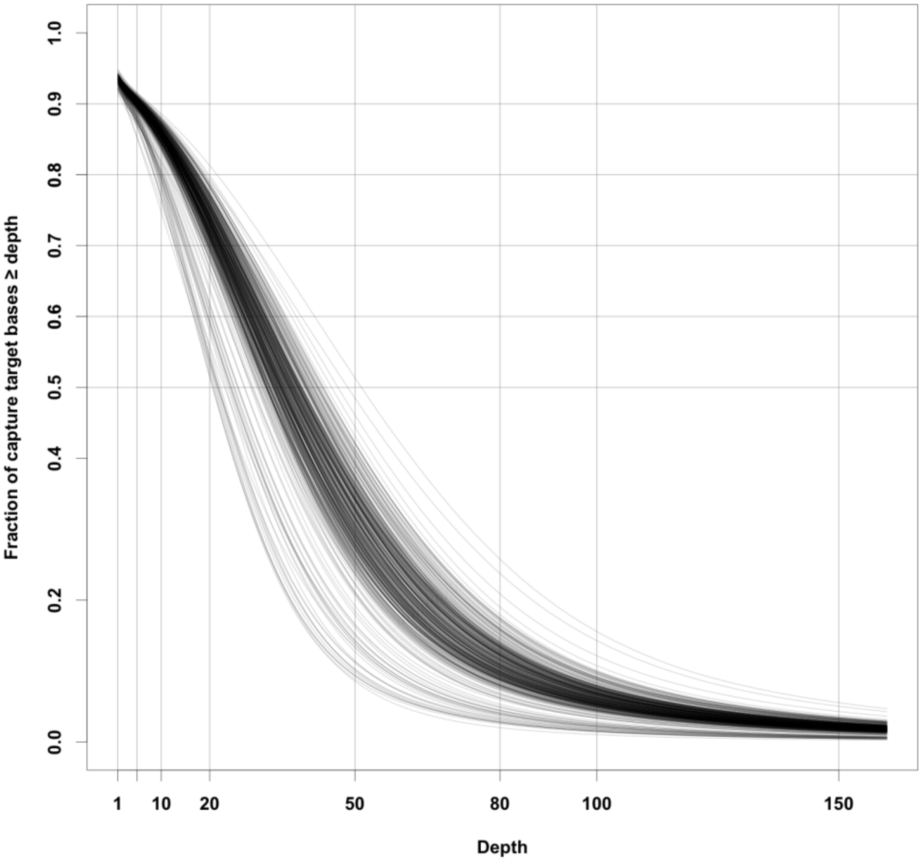

Fig. 2.3 Cumulative distribution of coverage depth of captured target bases in 375 trees. Each line represents a single tree

2.3.3 Single Nucleotide Polymorphism (SNP) Discovery

SNPs were detected in 375 individual trees using the SAMtools software v. 1.1 (Li et al. 2009). The raw SNPs were filtered using the selection criteria of being bi-allelic sites with at least 10X sequencing depth in at least 90 % of the individuals, and with the MAF ≥ 0.05. A total of 972,720 SNPs were acquired for downstream analyses.

These SNPs were located in 38,702 scaffolds of the loblolly pine reference genome assembly v. 1.01. A maximum of 854 SNPs were detected in one scaffold. Based on annotation of genomic regions, most of the identified SNPs resided in exons, but some resided in introns or unclassified regions. Among all the SNPs, 58 % were located in exons with an average SNP density of 11.5 SNPs/Kbp (one SNP per 87 bp); 53 % were located in CDS; 2 % in 5’ UTR; 3 % in 3’ UTR and 13 % in introns. By position relative to capture target region, 51 % of all SNPs were located in capture target regions with an average SNP density of 13.2 SNPs/Kbp (one SNP per 76 bp), and 49 % were located in off-target regions (Table 2.1). The number of SNPs detected in exons was more than in on-target regions because the capture extended to the adjacent area of each target.

One of the most important goals of exome sequencing is to identify the genetic variants that can be used in the association mapping analysis to dissect the phenotypes of interest. Such analyses require high quality SNPs, and therefore we focused only on those SNPs, both within and outside of exons, that passed the strict filtering criteria described above.

Table 2.1 Number and percent of 972,720 SNPs located in different genomic regions

Category	**SNPs**	**%**
Exon	564932	58.08
CDS	513652	52.81
5’ UTR	17693	1.81
3’ UTR	33587	3.45
Intron	127863	13.14
Unclassified	279925	28.78
On-target	498451	51.24
Off-target	474269	48.76

2.3.4 Population Genetics Metrics

SNPs with a MAF less than 0.05 were excluded, therefore SNP allele frequencies ranged between 0.05 and 0.5 with a median of 0.14 (Fig. 2.4). The average Ts/Tv ratio was 1.96 over all regions (Table 2.2 & 2.3). This value was higher in CDS than in UTRs. The transition bias could be attributed to natural selection on the nonsynonymous transversion, and the even higher ratio for CDS could be caused by the increased presence of methylated cytosine in CpG dinucleotides where the methylated cytosine can easily undergo deamination and transition to a thymine (Keller et al. 2007).

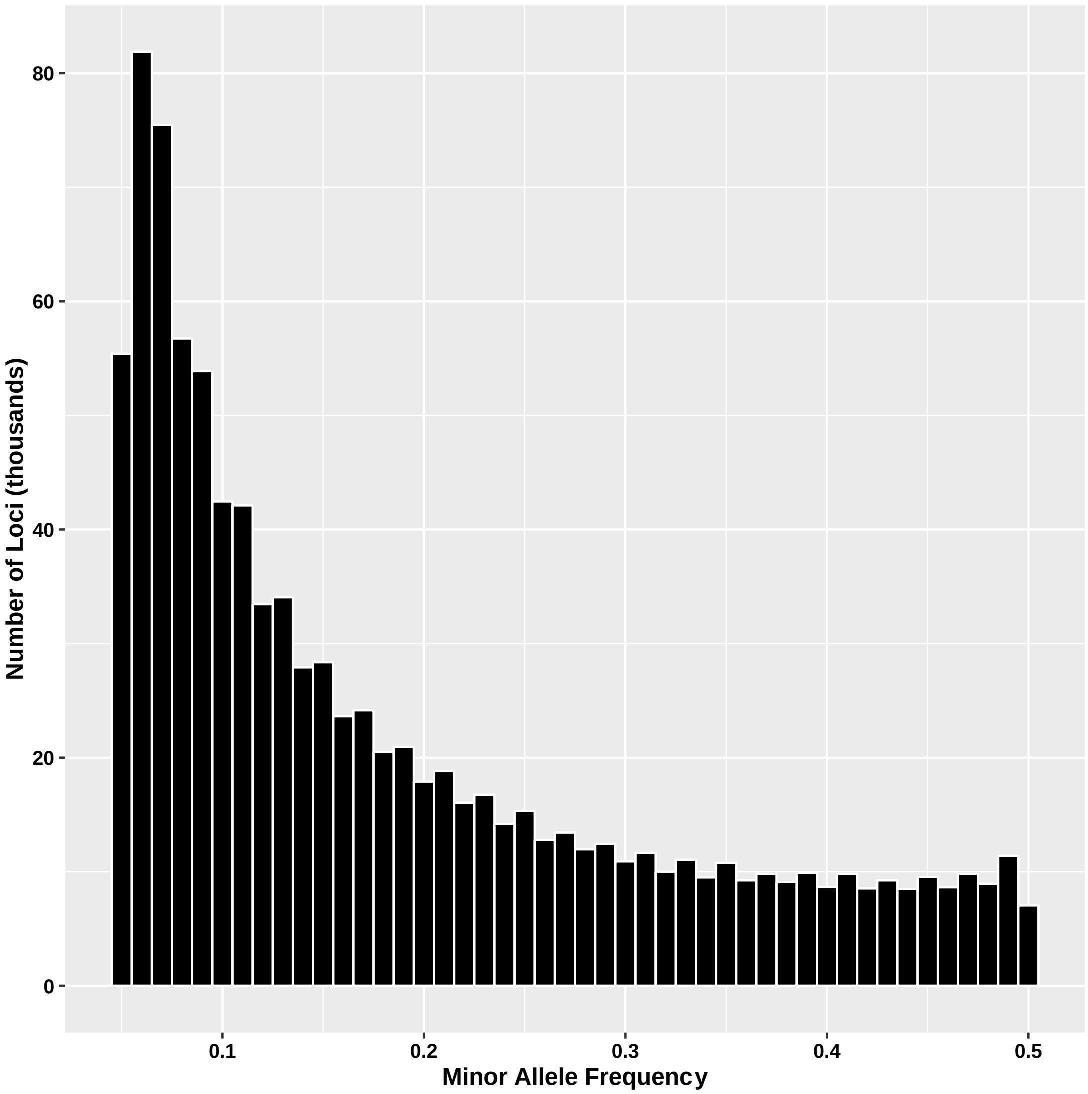

Fig. 2.4 MAF distribution among 972,720 SNPs

Table 2.2 Ts/Tv ratios for 972,720 SNPs categorized in different genomic regions

Total	**CDS**	**Exon**	**5’ UTR**	**3’ UTR**
1.96	1.98	1.93	1.58	1.45

Table 2.3 Transition (Ts) and transversion (Tv) nucleotide substitutions summary. Numbers of T_S and T_V for 972,720 SNPs in different genomic regions

Substitution type	**Total**	**CDS**	**Exon**	**5' UTR**	**3' UTR**
AC	89870	46655	52192	1834	3538
AG	320946	170211	186063	5388	9913
AT	63816	31523	36121	1341	3139
CG	85174	47396	52783	1871	3360
CT	322874	171247	187210	5438	9983
GT	90040	46620	52281	1821	3654
T_S	643820	341458	373273	10826	19896
T_V	328900	172194	193377	6867	13691

Heterozygosity and F_{IS} were estimated on an individual basis (Fig. 2.5). The results indicated a low inbreeding rate and a high level of genetic diversity. Among all individuals, the F_{IS} values were generally below zero, ranging between -0.24 and -0.06, except in tree 634A, where it was 0.21. Heterozygosity was between 0.29 and 0.33 except in 634A, where it was 0.21.

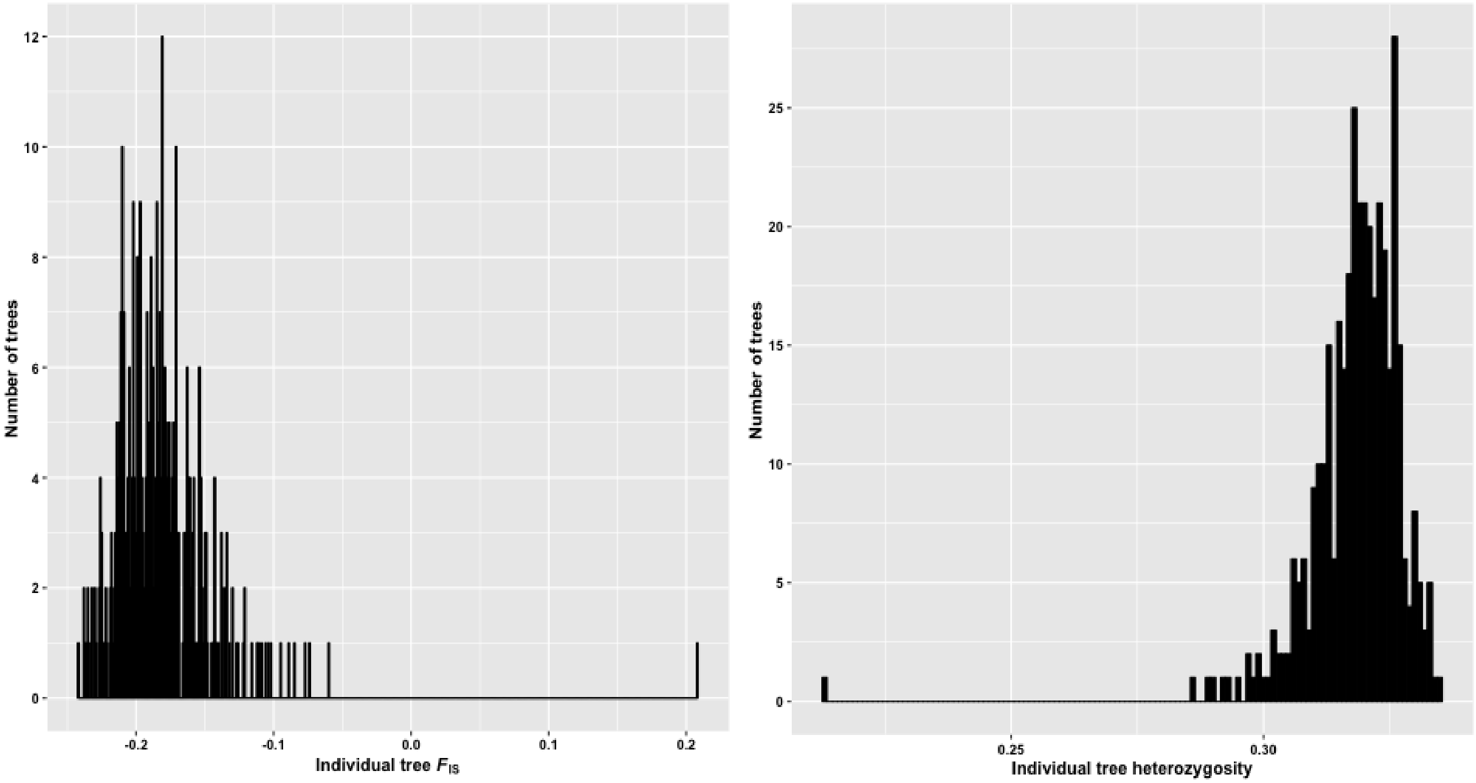

Fig. 2.5 F_{IS} (left) and heterozygosity (right) distributions among 375 trees

After Bonferroni correction (adjusted *P*-value < 5e-8), 188,072 (19 %) out of 972,720 SNPs significantly departed from Hardy-Weinberg equilibrium (HWE). Nucleotide diversity (π) in different genomic regions was estimated in a sliding window of 50 bp with a step of 25 bp (Table 2.4).

Table 2.4 Nucleotide diversity (π) in 50 bp step sliding windows for 972,720 SNPs in different genomic regions

Genomic region	Range	Mean	Median
Total	0.0016 – 0.1204	0.0119	0.0092
CDS	0.0016 – 0.1204	0.0117	0.0090
Exon	0.0016 – 0.1204	0.0116	0.0090
5’ UTR	0.0018 – 0.0787	0.0104	0.0083
3’ UTR	0.0017 – 0.0827	0.0100	0.0081
On Annotated Genes	0.0016 – 0.1204	0.0117	0.0090
Out of Annotated Genes	0.0016 – 0.1087	0.0122	0.0095

2.3.5 Genome-Wide Linkage Disequilibrium (LD)

LD is a non-random association of alleles at different loci and may indicate the genetic forces that structure the genome (Slatkin 2008). Investigations of genetic diversity and LD are prerequisites for association mapping and help in interpretation of results. We calculated the zygotic LD (squared correlation coefficient r^2) values for all SNP pairs within each scaffold in the genome assembly and plotted them against the physical distances between the same SNP pairs in the scaffold (Fig. 2.6). The average

LD for linked SNPs was inferred from the trendlines of the nonlinear regressions and started from 0.44, then decayed by half (0.22) at 55 bp, to 0.10 at 192 bp, and to 0.05 at 451 bp. The proportion of SNP pairs located within the same scaffold with r^2 > 0.1 was 18 % in this population, and with r^2 > 0.8 it was 3 %.

2.3.6 Population Structure

Evaluation of population structure is crucial for association mapping. If not accounted for, population structure may cause spurious associations between markers and phenotypes (Kang et al. 2008). The ADEPT2 population trees included in this study were the clonally-propagated, open-pollinated progeny of the originally sampled trees. The maternal origins were known for 362 out of 375 trees. The 362 trees can be divided into two sub-samples based on the geographic location of their maternal parents: 1) the sub-sample west of the Mississippi River represented by 55 trees from four states, and 2) the sub-sample east of the Mississippi River represented by 307 trees from eight states. F_{ST} was estimated on a per-site basis following Weir and Cockerham (Weir and Cockerham 1984). The F_{ST} range was between -0.01 and 0.72, with a median of 0.0087 (Fig. 2.7). The mean F_{ST} was 0.026, and the weighted F_{ST} was 0.028. Generally, the genetic differentiation between these two sub-samples was relatively low, but statistically significant.

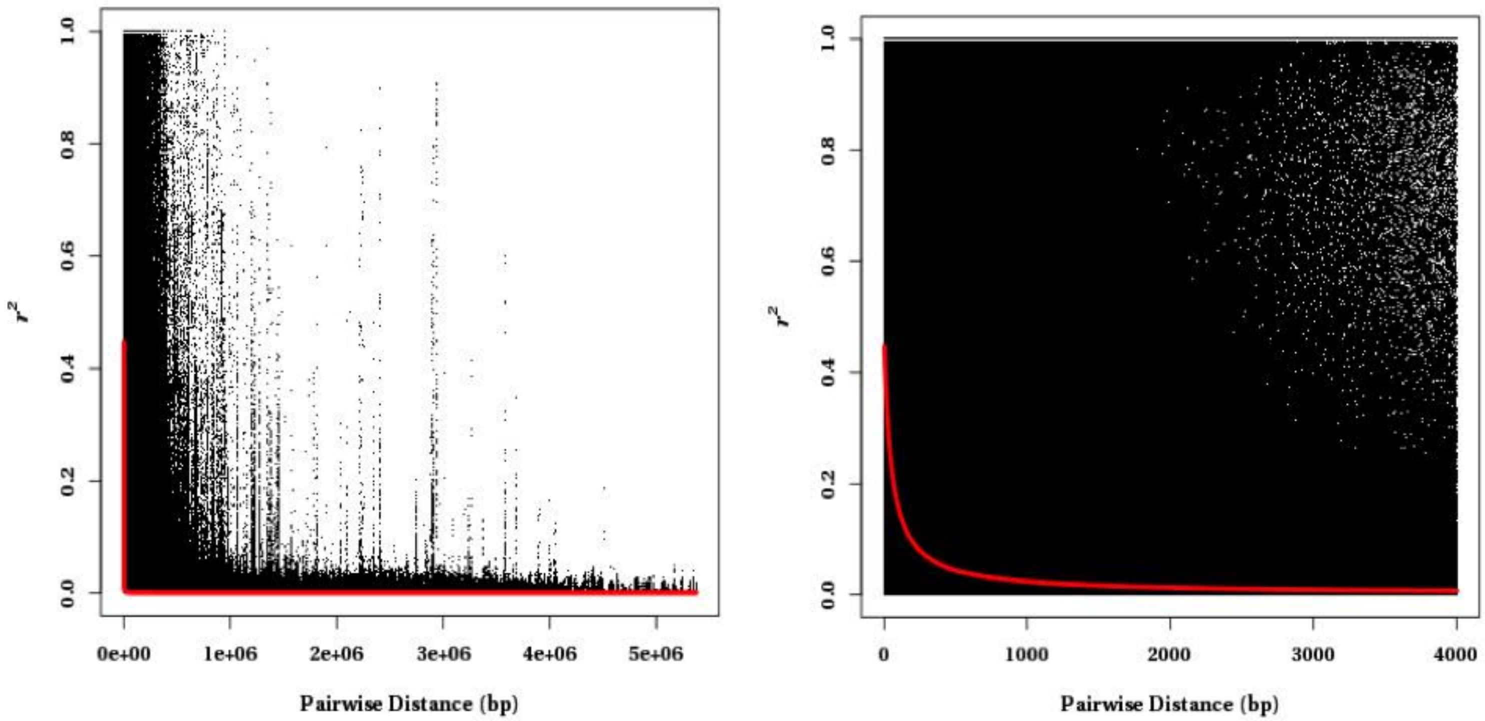

Fig. 2.6 LD decay plot for 375 trees based on 972,720 SNP markers. LD decay plot for 375 trees based on 972,720 SNP markers. Pairwise LD coefficients (r^2) calculated for all 375 trees were plotted against the physical distances (bp) between all pairs of SNPs within the same scaffolds (left) and between pairs of SNPs within the same scaffolds located within 4000 bp (right). The trendlines of the nonlinear regressions (r^2) against physical distance between the SNPs are indicated in red

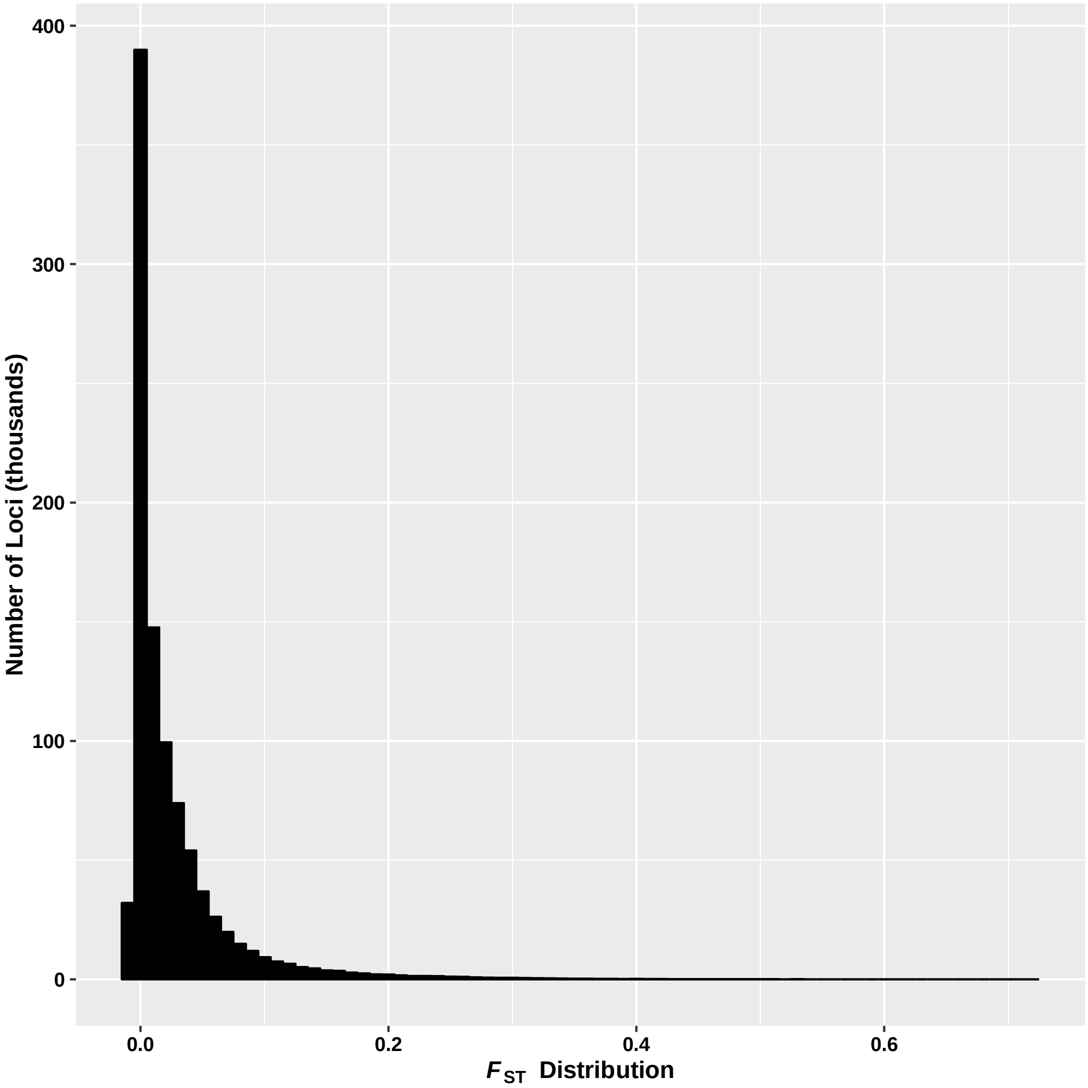

Fig. 2.7 F_{ST} distribution across all loci. The range is between -0.01 and 0.72, with a median of 0.0087. The mean F_{ST} is 0.026, and the weighted F_{ST} is 0.028

We then applied the software fastStructure (Raj et al. 2014) to infer the admixture proportion using our genotyping data. We thinned the marker set to no more than a single marker within 1 Mbp on each scaffold, which resulted in a presumably unlinked set of 30,146 SNPs. After testing a number of potential subpopulations (clusters) with fastStructure, ranging from $K = 1$ to $K = 12$ (where K is the number of subpopulations or clusters), we ran the recommended fastStructure algorithm for multiple K to choose the appropriate number of model components that explained structure in the dataset. The output showed model complexity that maximized marginal likelihood when $K = 2$, and the model components used to explain structure in data when $K = 7$. Therefore, we considered two and seven clusters as the most likely subpopulation clustering explaining the relationship between admixture proportion and geographical sites.

A clear geographical trend could be observed when the admixture proportions of each tree across clusters were plotted on a map (Fig. 2.8a & b). The segment in each pie chart corresponds to the summarized population assignment inferred by the software. We further aligned the admixture proportion of each tree with the longitude from west to east (Fig. 2.8c & d). Strong statistical correlations were observed between longitude and admixture proportion ($r^2 = 0.75$ when $K = 2$ and $r^2 = 0.74$ when $K = 7$). In Fig. 2.8c & d, vertical lines arranged from left to right correspond to the individual trees according to their original maternal parents' geographic locations from west (Texas) to east (North Carolina) in the southeastern U.S. Each vertical line represents admixture proportions for an individual tree partitioned when $K = 2$ (Fig. 2.8c) or $K = 7$ (Fig. 2.8d). The left 55 trees on the X-axis represent the trees west of the Mississippi River, while the other trees are from east of the Mississippi River.

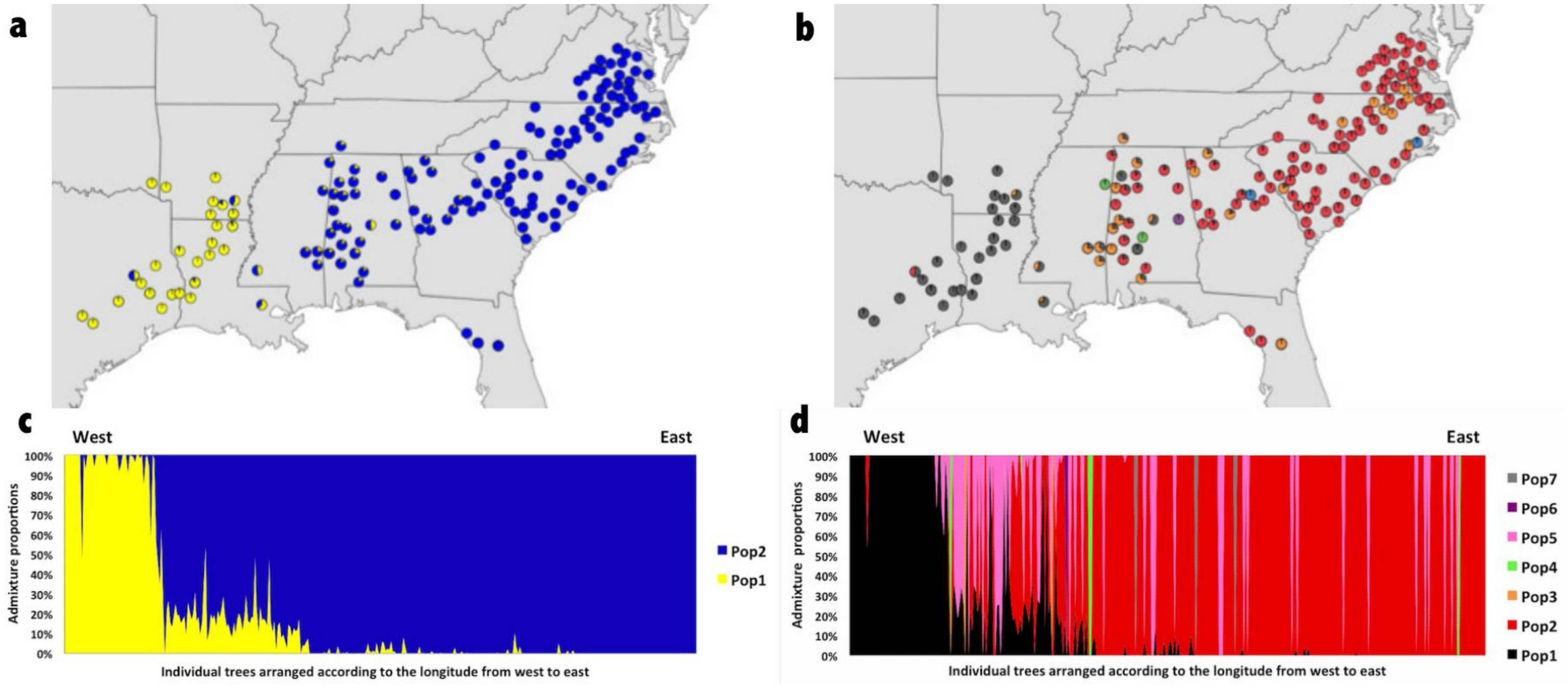

Fig. 2.8 Summarized admixture proportion distributions for $K = 2$ and $K = 7$. a & b Summarized admixture proportions plotted on the map. Each pie chart is partitioned via summarized population assignments inferred by fastStructure. c & d Individual tree admixture proportion distributions. The trees are aligned on the x-axis according to the longitude from west to east

2.4 Discussion

In the first published study of exome capture in loblolly pine, 54,773 probes representing 6.57 Mbp of target exome were designed using 14,729 unique transcripts derived from the assembly of ESTs (Neves et al. 2013b). However, the unavailability of a reference genome and, therefore, lack of information on the exon-intron boundaries, negatively affected the probe design. This caused insufficient capture and cross-hybridization and decreased the capture efficiency. This problem was mitigated in our exome capture study, because the probe set covered almost the entire exome and its design took into account the exon-intron structure. The designed probe set covered ~46 Mbp of target exome and included previously uninvestigated genomic regions. The risk of capturing pseudogenes was decreased by using only genes classified as “high quality” to design the probes. A key concern during the probe design was the exclusion of those probes that might cross-hybridize with non-target regions and repetitive elements, especially considering that 82 % of the loblolly pine genome consists of the highly repetitive sequences (Neale et al. 2014). In this study, the preliminary probes were stringently filtered to exclude possible cross-hybridization with non-target regions and repetitive elements. Although the capture size could be potentially expanded, if the filtering criteria had been relaxed, the stringent filter guaranteed the hybridization specificity and prevented cross-hybridization.

Multiplexing individually and uniquely indexed samples before capturing and sequencing greatly saves time and money and has become a standard procedure in sequence capture experiments. However, sufficient sequencing depth (output) is still needed to guarantee a higher coverage depth on the target regions. Fig. 2.2b & c demonstrate that the coverage depth is positively correlated with the sequencing output. Therefore, multiplexing should be reasonable and should ensure sufficient on-target coverage depth to avoid problems associated with low SNP detection power. In our study, uneven numbers of sequencing reads across different individual tree samples could be mainly due to multiplexing of unequal amounts of the sample libraries.

Some of the reads could not be mapped to the reference genome, likely due to either incomplete assembly of the reference genome or multiple sequencing errors in the reads that exceeded the mismatch tolerance threshold of the mapping parameters. Although the probes were filtered for cross-hybridization prior to the actual hybridization step, further filtering of the multi- and improperly mapped reads was important in order to retain only the high quality mapped reads for downstream analyses. Similarly, the redundant reads were also filtered to remove the potential PCR duplicates and to correct the coverage depth.

The read mapping results demonstrate a high level of on-target efficiency in this research. This guarantees the target regions have enough coverage depth. Less than 9% of the target regions had no matching reads. The main reason for this was that the probes covering these regions were filtered out to avoid cross-hybridization. It should also be noted that the current reference genome assembly is still under development and the target regions with no matching reads could potentially be artifacts or mis-assembled parts of the reference genome.

The high level of genetic diversity was expected because loblolly pine is a highly outcrossing and polymorphic species. In addition, the ADEPT2 population was established for association mapping with presumably unrelated trees originally sampled from across a wide part of the natural range. Tree 634A may be a progeny from selfing or a mating between closely related trees.

Regions out of annotated genes had higher average nucleotide diversity than in annotated genes. This could be due to selection constraints. However, it should be noted that the highly diverged sequences could not map to the reference genome, hence biasing the diversity estimates.

Highly outcrossing conifers are expected to have a rapid LD decay. Neale and Savolainen (Neale and Savolainen 2004) reported that the r^2 decayed to less than 0.20 within ~1500 bp based on 19 candidate genes in loblolly pine. In spruces, LD displayed diverse patterns among different genes or the same genes in different species, declining rapidly to half between a few base pairs and 2000 bp (Namroud et al. 2010). In Douglas-fir (*Pseudotsuga menziesii*), LD decayed > 50 % over relatively short segments from r^2 = 0.25 to 0.10 within 2000 bp based on sequencing 18 genes (Krutovsky and Neale 2005). LD estimates in this study based on the exome-derived sequences indicated an even faster decay than previously reported. This could be due to the much larger number of gene regions analyzed in this study. The discrepancies can be partly explained also by different methods used for estimating LD. The abovementioned studies calculated gametic LD statistics r^2 using megagametophyte haplotypes, while in this study, zygotic LD between genotypes was calculated. However, gametic LD can also be calculated in our study based on the inferred (phased) haplotypes. When we used the phased haplotypes inferred by the software Beagle v. 4.1 (Browning and Browning 2007) for the 972,720 SNPs to calculate gametic LD, a slower decay was observed, with LD decaying by half (r^2 = 0.22) at 79 bp and to r^2 = 0.10 at 280 bp. The rate of LD decay can vary between genes and across different genome regions (Namroud et al. 2010). Therefore the generality of LD distribution across the entire loblolly pine genome remains to be further analyzed because only a relatively small and highly specific part of the entire genome was studied here. Our study relied also on the accuracy of contig and scaffold assembly in the draft reference genome that should be verified and ordered in the future studies.

It has been widely recognized that the glacial advance and retreat have altered the landscape of the Mississippi Valley and the species became restricted into glacial refugia, thus high dissimilarity was formed between refugia populations (Pessino et al. 2014). A postglacial barrier to dispersal was created between populations located on west and east Mississippi River and thus decreased the gene exchange and increased the overall genetic variance in some species (Maggs et al. 2008; Pessino et al. 2014). The discontinuity is also evident in loblolly pine, as can be concluded from genetic differentiation estimated in our study based on ADEPT2 population, and in the earlier studies that were based on limited numbers of SNP and SSR markers (Al-Rabab'ah and Williams 2002; Eckert et al. 2010).

3. ASSOCIATION GENETICS OF QUANTITATIVE TRAITS

3.1 Introduction

Advanced loblolly pine breeding practice has been implemented over the past 50 years, creating favorable production economics (McKeand et al. 2006). Appropriate breeding strategies rely on an understanding of valuable traits including crown structural characteristics, growth, water use efficiency (WUE) and disease resistance. Crown structural characteristics such as branch angle, leaf area and crown width affect interception of radiation and competition with other trees (Emhart et al. 2007). Larger trees tend to display flatter angles and have wider crowns and thus expose more leaf area (Lambeth and Hubert 1997; Emhart et al. 2007). The ability to sustain yield and quality under adverse conditions such as drought and diseases is also a key consideration for loblolly pine breeding. Carbon isotope discrimination ($\Delta^{13}C$) has long been used to reflect long-term WUE in forest trees due to the feasibility to screen a large number of individuals over a short period of time (Aitken et al. 1995; Baltunis et al. 2008). Plants with higher WUE discriminate less against ^{13}C when they are exposed to the same fluctuations in environmental conditions (Farquhar et al. 1989; Aitken et al. 1995; Cregg and Zhang 2000; Baltunis et al. 2008; Cumbie et al. 2011). Pitch canker disease in southern pines caused by the fungus *Fusarium circinatum* has been a serious problem in the last a couple of decades. It endangers loblolly pine via the resinous lesions on stems and branches that lead to high seedling mortality and slower growth rates (Kayihan et al. 2005; Quesada et al. 2010).

Association mapping can be used to dissect these traits of interest, identifying genes controlling them. Association studies require a mapping population representing a wide spectrum of phenotypic variation, phenotypic measurements and abundant molecular markers. Environmental conditions drive forest phenotypic adaptation and geographic distribution through natural selection. Within the loblolly pine species, trees from different provenances have considerable diversity in phenotypic performance (Schmidtling 2001). To grasp as much phenotypic variation as possible for our association study we selected the loblolly pine clonally propagated population from the Allele Discovery of Economic Traits in Pine 2 (ADEPT2) project. It was specifically designed for association mapping to represent a range-wide collection from regions with different environmental conditions (see Plant material in Materials and methods below for details). An array of adaptive traits, namely, specific leaf area, branch angle, crown width, stem diameter, height, Δ^{13}C and nitrogen concentration were measured in this population.

Loblolly pine occurs naturally on both sides of the Mississippi River, and there are differences observed between eastern and western sources. This presumably can be dated back to the Pleistocene geologic era. During the last glaciation, loblolly pine retreated to two refugia, one in southeast Texas and/or northeast Mexico, and the other in south Florida and the Caribbean. Isolation prevented pollen dispersal between populations and hence resulted in some of the differences observed now (Wells et al. 1991; Schmidtling 2001; Eckert et al. 2010). The population structure should be taken into account in association studies.

The abundance of genetic variation in the abovementioned traits and their value for adaptation or growth suggest they are potential subjects for artificial selection. With advances in molecular and genomics methods, the selection process may be accelerated in forest trees breeding by marker-assisted selection and genomic selection (Thavamanikumar et al. 2013). However, these methods have not been widely applied in the breeding of loblolly pines or other conifers. One reason is that the large and complex genome of loblolly pines poses challenges for gene discovery (Neale et al. 2014; Wegrzyn et al. 2014), association studies and genomic selection (Resende et al. 2012; Isik 2014). The number and identity of the genes controlling productivity and adaptive traits are largely unknown (Gonzalez-Martinez et al. 2006). Secondly, high-density SNP genotyping assays/platforms remain to be developed although 7K and 9K Illumina Infinium SNP genotyping arrays for loblolly (Eckert et al. 2009b) and maritime (Plomion et al. 2016) pines, respectively, are available. Due to these problems and a long generation time, only traditional family based breeding methods are currently applied to these forest trees.

Association analysis is an efficient method to dissect complex traits using SNPs. SNPs are abundant in the loblolly pine genome and their statistical association with phenotypes can be detected using association mapping, thus identifying the genes and their effects that underlie complex traits (Gonzalez-Martinez et al. 2006). Successful association mapping has been implemented within diverse loblolly pine populations and produced an array of SNP markers and genes that were connected with various traits. Using nearly 4000 SNPs derived from amplicons of 18 megagametophytes with PCR primers designed from unique Expressed Sequence Tag (EST) based contigs (Eckert et al. 2009a), 10 SNPs were detected to be associated with pitch canker disease resistance (Quesada et al. 2010), 7 SNPs with $\Delta^{13}C$, 1 SNP with height, 6 SNPs with foliar nitrogen concentration (Cumbie et al. 2011), 28 SNPs with metabolites (Eckert et al. 2012), and 80 SNPs with expression of 33 xylem development genes (Palle et al. 2013). Additionally, 101 associations with 27 gene expression phenotypes (Seeve 2010), and numerous SNPs with height, diameter at breast height, volume, fusiform rust resistance, wood specific gravity and stem forking index (Chhatre et al. 2013) were identified.

Problems and challenges have emerged with the application of association mapping to loblolly pines and other conifer species. First, for most quantitative and complex traits, large numbers of alleles explain only a small portion of the genetically heritable variation (Quesada et al. 2010; Cumbie et al. 2011). The missing heritability might be due to the small number of markers and rare variants that were usually excluded from the genotyping chips (Manolio et al. 2009). Second, though most efforts were put on the discovery of additive marker associations, non-additive effects imposed by dominance and epistasis play important roles in determining the genetic variation and need to be further studied (Eckert et al. 2009a; Cumbie et al. 2011). Third, the gene-based SNP discovery focused primarily on coding sequence (CDS) regions. However, regulatory elements in non-coding regions tend to have more polymorphisms with small effects associated with quantitative traits (Flint and Mackay 2009). Fourth, epigenetic modifications control most phenotypic variation (Cortijo et al. 2014; Kawakatsu et al. 2016), especially in forest trees that have extensive epigenetic modifications and phenotypic plasticity (Fabbrini et al. 2012; Yakovlev et al. 2012; Brautigam et al. 2013; Benomar et al. 2016; Gugger et al. 2016). For instance, C-effects due to clonal propagation via rooted cuttings are related to epigenetic modifications and confound with phenotypic differences.

The key steps to addressing the aforementioned problems are genetic variation discovery followed by association analyses. Target enrichment combined with genotyping by sequencing (GBS) technologies provides opportunities to survey large-scale populations for ample variants including rare alleles in a cost effective and efficient manner (Neves et al. 2013b; Suren et al. 2016). We utilized the NimbleGen SeqCap EZ system (Roche NimbleGen, Inc., Madison, WI) for genome target enrichment to discover over 2.8 million SNPs using exon-based probes and the DNA of 375 trees from a clonally propagated loblolly pine population (Lu et al. 2016). Phenotypic data collected from this population was associated with the genotyped SNPs to investigate the associations of genetic variation with the traits. The genetic correlations between traits, the geographical variation within the traits and exome-wide individual heterozygosity-trait correlations (HTC) were also examined. The objective of this study was to explore the genetic factors that influence adaptive performance of loblolly pines and to contribute to future breeding efforts.

3.2 Materials and Methods

3.2.1 Plant Material

The loblolly pine population used in this study was originally established for the Allele Discovery of Economic Pine Traits 2 (ADEPT2) project (Cumbie et al. 2011). Maternal parents of the ADEPT2 population were sampled across the natural range of loblolly pine. Trees were grown from open-pollinated seeds for one year and then were hedged and established for use in the ADEPT2 project. During the spring of 2010, rooted cuttings of 384 trees from the ADEPT2 population were established at the Harrison Experimental Forest at the Southern Institute of Forest Genetics (30°63' N, 89°06' W, Saucier, Mississippi). A randomized incomplete block alpha design was used, with 3 replications of 24 incomplete blocks of size 16 ($r = 3$, $s = 24$, $k = 16$, 4 trees × 4 trees in each block). These trees were used for phenotyping and collection of foliage for DNA isolation, exome enrichment and genotyping by sequencing.

3.2.2 Phenotyping

Sample collection and measurement of most traits of interest were conducted during the fourth growing season at the Harrison Experimental Forest. Between May 25 and June 24, 2014, the following traits were measured and recorded for each tree: total height was measured using a meter pole; branch angle, represented by the average of three branch angles relative to level at the third major whorl from the top, was measured using a digital level inclinometer; stem diameter at 18 inches high above the ground was measured using a caliper; crown width at the third major whorl from the top was measured using a measuring tape. In addition, total height was measured again in 2015 before and after the growing season, plus, height growth in 2015 was calculated as the difference between total height before and after the growing season.

South-facing and fully expanded needles from a point half way to the top of each tree were collected for assessment of specific leaf area, $\Delta^{13}C$ and nitrogen concentration. Specific leaf area was calculated as 20 needles' leaf area divided by the dry weight of these 20 needles. The leaf area was measured using a LAI 3000 scanner (Li-Cor, Lincoln, NE). The needles were dried at 65 °C for 72 hours.

For $\Delta^{13}C$ and nitrogen concentration analyses, 5 or 6 needles from each dried needle sample were ground into fine, homogeneous powders with a ball mill (MM400, Retsch, Hann, Germany). The samples were weighed in tin capsules and analyzed by EA-IRMS (Delta V, Thermo Scientific, Waltham, MA) in the Stable Isotopes for Biosphere Science Laboratory at Texas A&M University (http://sibs.tamu.edu; College Station, TX). The carbon isotope ratios were reported against VPDB with calibrated laboratory standards. We calculated $\Delta^{13}C$ values using the formula $\frac{\delta a - \delta p}{1+\delta p}$ (Farquhar et al. 1989), where δa and δp represented the isotope composition of air and leaf tissue, respectively (δa was assumed to be -8 ‰). The nitrogen concentration was reported as a mass percentage.

The pitch canker disease resistance data was taken from a published study conducted at the University of Florida (Gainesville, FL) (Quesada et al. 2010). The mean lesion lengths from 4 replications per clone were available for 317 trees used in this study. Therefore, only 317 trees were included for pitch canker resistance analyses.

3.2.3 Phenotypic Data Analyses

A mixed model analysis (Mclean et al. 1991) was used in order to assess the clonal effects for the measured traits: $y_{ijk} = \mu+r_i+b_k(r_i)+c_j+r_ic_j+e_{ijk}$, where y_{ijk} is the phenotypic value for the *j*th clone in the *i*th replication and *k*th block, μ is the population mean, r_i is the fixed variable of replication (i=1-3), c_j is the random variable of clone (j = 1-384, approx. NID(0, σ^2_C)), $b_k(r_i)$ is the random variable of block nested within replication (k = 1-24, approx. NID(0, $\sigma^2_{b(r)}$)), r_ic_j is the random variable for the interaction of replication by clone (approx. NID(0, σ^2_{rc})), and e_{ijk} is the error term (approx. NID(0, σ^2_e)). The best linear unbiased prediction (BLUP) estimates for each trait were used as phenotypic values in further analyses. The clonal repeatability was estimated using the formula: $r_{clone}= \frac{\sigma^2_{clone}}{\sigma^2_{clone}+\frac{\sigma^2_e}{rep}}$

The BLUP estimate for mean lesion length induced by pitch canker disease was acquired using the model: $y_{ij} = \mu+r_i+c_j+r_ic_j+e_{ijk}$, where y_{ij} is the phenotypic value for the *j*th clone in the *i*th replication, μ is the population mean, r_i is the fixed variable of replication (i=1-4), c_j is the random variable of clone (j=1-317, approx. NID(0, σ^2_C)), r_ic_j is the random variable for the interaction of replication by clone (approx. NID(0, σ^2_{rc})), and e_{ij} is the error term (approx. NID(0, σ^2_e)).

A total of 362 trees within this population have known maternal origins. This population was divided into 3 regions as described by Schmidtling (2001) using maternal origins. The eastern region includes states east of the Mississippi River, the western region includes the states of Arkansas and Louisiana, and the far west region includes the states of Texas and Oklahoma (Fig. 1). An analysis of variance (ANOVA) was applied to compare the BLUP differences of each trait for individuals grouped by their regions of maternal origin. All the statistical analyses were conducted using the JMP Pro 12 statistical software (SAS Institute, Cary, NC).

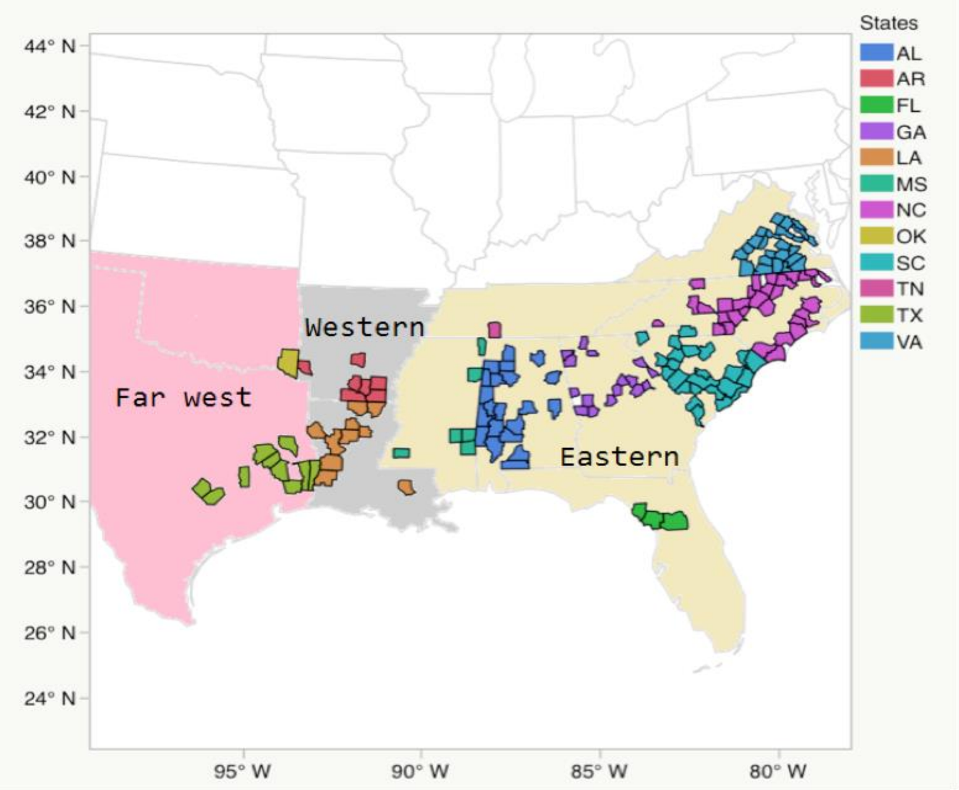

Fig. 3.1 The counties of origin of the studied maternal loblolly pine trees. The range was divided into 3 regions, Far west (the states of Texas and Oklahoma, highlighted by pink color), Western (the states of Arkansas and Louisiana, highlighted by grey color) and Eastern (east of the Mississippi River, highlighted by beige color)

3.2.4 Genotypic Data

Genotypic data were obtained by the authors for 375 trees in this ADEPT2 population (Lu et al. 2016). The NimbleGen SeqCap EZ system (Roche NimbleGen, Inc., Madison, WI) was used to capture and enrich the exome of each tree. The detailed procedures of probe design, raw SNP detection and genotyping are described in Lu et al. (2016). In this study, the raw SNPs were filtered, accepting only bi-allelic sites with at least 5X sequencing depth for all of the individuals without missing data and a minor allele frequency (MAF) ≥ 0.01. A total of 2,822,609 SNPs were retained. Among these SNPs, 1,199,938 (43 %) reside in CDS, 36,533 (1 %) in five prime untranslated regions (5’ UTR), 70,377 (2 %) in three prime untranslated regions (3’ UTR), 516,268 (18 %) in introns and the remaining SNPs (36 %) in unclassified regions, probably unannotated regulatory elements of genes or intergenic regions. A total of 94,478 haplotype blocks were detected for this population using PLINK 1.9 (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al. 2007).

3.2.5 Association Analyses

Association analyses were conducted using TASSEL 5.0 (Bradbury et al. 2007). SSR markers that could be used for estimating covariates to adjust for population structure were previously genotyped in 249 out of the 375 genotyped trees in this population. We identified these 249 trees as an individual population, named the structure (*str*) population. Population structure within this population was mainly due to the Mississippi River discontinuity (Lu et al. 2016). We identified the 307 trees from east of the Mississippi River as an individual population named the *east* population. The three populations: *total*, *east* and *str* populations, were used to perform association analyses. For the *total* and *east* populations, the simple general linear model (GLM) method (*S* model) and the mixed linear model (MLM) method incorporating a kinship matrix (*K* model) were applied. For the *str* population, in addition to the *S* and *K* models, the GLM incorporating the covariate to adjust for population structure (*Q* model) and the MLM incorporating both the kinship matrix and population structure covariate (*QK* model) were applied. Population structure covariate was estimated using the software STRUCTURE (Pritchard et al. 2000; Hubisz et al. 2009) and 23 SSR markers as described by Eckert et al. (Eckert et al. 2010). A kinship matrix for each population was estimated by TASSEL 5.0 (Bradbury et al. 2007) using the SNP markers. Quantile-quantile plots were generated for observed against expected $–\log_{10}(P)$ to examine the model fitness, where observed *P*-values were obtained from association mapping and expected *P*-values from the assumption that no association occurred between marker and trait. Significance of associations between loci and traits were determined by the *P*-values. The Bonferroni threshold was 0.05/2,822,609=1.77E-8, where 2,822,609 was the number of total SNPs. However, this threshold was overly conservative. Instead, a corrected Bonferroni threshold 0.05/94,478=5.29E-7, where 94,478 was the number of haplotype blocks, was applied to screen for significant loci.

3.2.6 Individual Exome-Wide Heterozygosity-Trait Correlation (HTC) Analyses

The observed individual exome-wide heterozygosity values were calculated for each of 375 trees within this population using the 2,822,609 genotyped SNPs and the software VCFtools (Danecek et al. 2011). BLUP estimates for each trait were used to correlate with the heterozygosity (HTCs). Pearson correlation coefficients (*r*) were used to evaluate the HTCs. Within this population, 362 trees have known maternal origins and could be separated accordingly to the two populations on the west and east sides of the Mississippi River, respectively. Since distinct geographical and genetic structure patterns along the Mississippi River exist within this population, *r* was calculated separately in the *total* (*n* = 375), *east* (*n* = 307) and *west* (*n* = 55) populations.

3.2.7 SNP Interaction Analyses

The epistatic SNP-SNP interaction test was implemented using PLINK 1.9 (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al. 2007). A Bonferroni threshold of 0.05/3,176,320,469,273=1.57E-14, where 3,176,320,469,273 was the number of all SNP pairs, was used to correct the multiple comparisons.

3.2.8 Annotation of Genes that Contained SNPs Associated with Traits

The information for genes that contained SNPs associated with traits was obtained from loblolly pine Gene Annotation v3.0 (http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/Pita_Annotation_v3.0/) (Wegrzyn et al. 2014). The loblolly pine reference genome assembly and annotation are under active improvement. The regulatory sequences such as promoters, enhancers and silencers have not been identified yet. SNPs within 5000 bp downstream or upstream of a gene were considered to be within a putative regulatory sequence of the gene. If a SNP was located in a region without annotation, the flanking sequence 700 bp upstream and downstream of the SNP was used as a query to do a Blastx search against the entire National Center for Biotechnology Information (NCBI) nonredundant (nr) protein database (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

The percentage of clonal (clonal values were obtained using BLUP) and phenotypic variance explained by the identified SNPs or SNP-SNP interactions were estimated by comparing the model incorporating the SNPs or SNP-SNP interactions as random effects with the reduced model without SNP effects. To obtain the additive and dominance effects for the SNPs detected by association analyses, the loci in Hardy Weinberg Equilibrium with all three genotype classes present were treated as fixed effects and tested in linear regressions.

For the identified SNPs located in CDS, the effect of SNP substitution on the amino acid was investigated by aligning the sequences with the SNPs and the corresponding transcripts from the loblolly pine Gene Annotation v3.0 (http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/Pita_Annotation_v3.0/) (Wegrzyn et al. 2014) using the Clustal Omega software (http://www.ebi.ac.uk/Tools/msa/clustalo/) (Sievers et al. 2011). The software ExPASy Translate (http://web.expasy.org/translate/) (Gasteiger et al. 2003) was utilized to translate the DNA sequences to amino acid sequences.

3.3 Results

3.3.1 Clonal Repeatability, Genetic Correlations and Geographical Variation

Clonal loblolly pines from the ADEPT2 population were measured for traits of interest during their fourth growing season at the Harrison Experimental Forest. At the time of phenotyping, 78 % of the trees planted in 2010 had survived. For each trait measured, we calculated the clonal repeatability within this population (Table 3.1). The clonal repeatability value is an overestimate of narrow-sense heritability (Havens 1994), it is also a conservative estimate of broad-sense heritability. Clonal repeatability calculations indicated that 31 % of the variation in specific leaf area could be attributed to genetic effects. For growth and architecture traits, 60 % of the variation in branch angle, 62 % in crown width, 56-62 % in total height, 54 % in stem diameter and 11 % in 2015 height growth could be attributed to genetic effects. Due to the low clonal repeatability observed, height growth in 2015 was excluded from further analyses. Clonal repeatability values for Δ^{13}C (85 %) and nitrogen concentration (76 %) were higher than for other traits.

Bivariate Pearson Correlation was conducted to investigate the genetic correlations between the traits (Table 3.2). Strong positive correlations were observed between total height and crown width, diameter and crown width, and diameter and total height. Δ^{13}C was correlated with branch angle, crown width, total height and nitrogen concentration. A small negative correlation existed between crown width and branch angle. Nitrogen concentration was positively correlated with specific leaf area, crown width and total height.

Table 3.1 Phenotypic data summary

Trait	Mean	Standard deviation	Maximum	Median	Minimum	Number	Clonal repeatability
SLA[a], cm^2/mg	28.88	5.19	43.38	29.06	14.11	920	0.31
BA[b], °	35.36	8.75	62.50	35.93	9.30	921	0.60
CW[c], m	1.40	0.34	3.60	1.39	0.43	922	0.62
DIA[d], m	0.05	0.02	0.10	0.05	0.01	922	0.54
2014H[e], m	3.25	0.73	5.44	3.30	0.85	922	0.56
2015HB[f], m	3.99	0.93	6.78	4.04	1.07	918	0.61
2015HA [g], m	4.66	0.97	7.44	4.70	1.52	908	0.62
2015HG[h], m	0.67	0.23	2.41	0.65	0.07	908	0.11
$\Delta^{13}C$ [i], permil, VPDB	24.28	0.56	25.81	24.30	22.21	920	0.85
N[j], %	0.93	0.09	1.28	0.94	0.58	920	0.76

[a] Specific leaf area; [b] Branch angle; [c] Crown width; [d] Stem diameter; [e] Total height in 2014; [f] Total height in 2015 before the growing season; [g] Total height in 2015 after the growing season; [h] Height growth in 2015; [i] Carbon isotope discrimination; [j] Nitrogen concentration

We compared BLUP estimates of individuals grouped by their regions of maternal origin (Fig. 3.1). Significant differences ($P < 0.05$) were identified for crown width ($P = 0.002$), stem diameter ($P = 0.005$), nitrogen concentration ($P = 0.044$) and $\Delta^{13}C$ ($P = 0.004$) between eastern, western and far west regions (Fig. 3.2). For the other traits, no significant differences were found among different regions. Trees from the eastern region showed greater crown width, stem diameter and nitrogen concentration than those from the western and far west regions, indicating genotypes from east of the Mississippi River tend to have a higher growth rate than those from west. The lower $\Delta^{13}C$ values of eastern trees indicate higher WUE than that of trees from western and far west regions.

Table 3.2 Pearson correlation coefficients between the traits

	SLA[a]	**BA**[b]	**CW**[c]	**2014H**[e]	**DIA**[d]	**$\Delta^{13}C$**[i]	**N**[j]
SLA[a]		[h]NS	NS	NS	NS	NS	0.140
BA[b]			-0.107	NS	NS	0.154	NS
CW[c]				0.757	0.769	-0.189	0.123
2014H[e]					0.897	-0.138	0.113
DIA[d]						NS	NS
$\Delta^{13}C$[i]							-0.251
N[j]							

[a] Specific leaf area; [b] Branch angle; [c] Crown width; [d] Stem diameter; [e] Total height in 2014; [i] Carbon isotope discrimination; [j] Nitrogen concentration; [h]Non-significant at $P < 0.05$

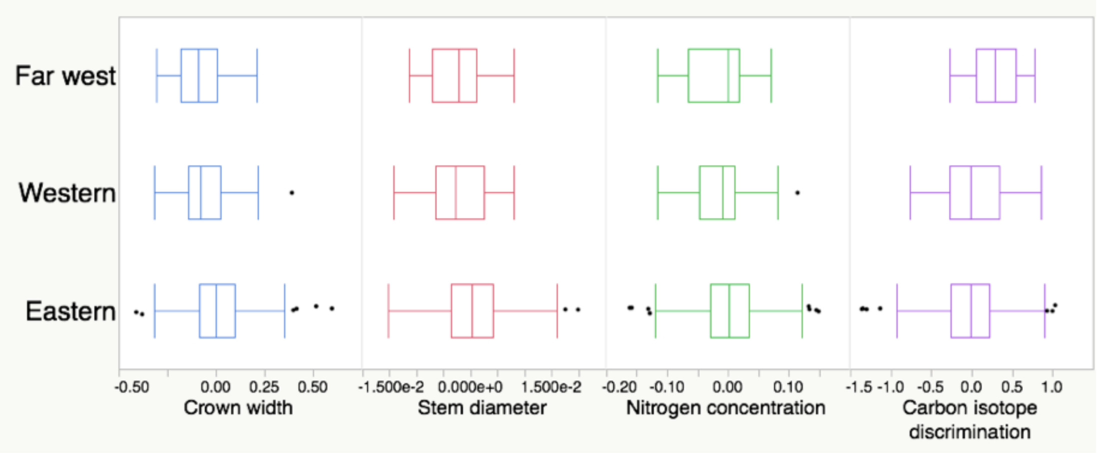

Fig. 3.2 BLUP estimates distribution for the traits with significant differences (*P* < 0.05) among Far-west, Western and Eastern regions

3.3.2 Individual Exome-Wide HTCs

To test whether gene-based individual exome-wide heterozygosity affects adaptive traits, we used exome-based SNPs to calculate individual multi-locus heterozygosity and correlated the heterozygosity values with the BLUP estimates of each trait within the clonally tested populations. Two significant HTCs were detected in the *total* population, and one HTC in the *east* population (Table 3.3). No significant correlations were detected in the *west* population. Heterozygosity was negatively correlated with $\Delta^{13}C$ (r = -0.173 in the *total* population and -0.137 in the *east* population). Since a smaller $\Delta^{13}C$ value indicates higher WUE, the negative relationship between the heterozygosity and the $\Delta^{13}C$ may indicate that higher heterozygosity is positively associated with WUE. The positive correlation between the heterozygosity and the nitrogen concentration (*r* = 0.124) probably suggests heterozygosity is also positively associated with nitrogen concentration.

Table 3.3 Significant HTCs

Population (number of trees)	**Trait**	***r***	***p***
Total (375)	Carbon isotope discrimination	-0.173	0.001
	Nitrogen concentration	0.124	0.017
East (307)	Carbon isotope discrimination	-0.137	0.017

r , Person correlation coefficient

3.3.3 Marker-Trait Association Analyses

Association analyses were performed using *S* and *K* models for the *total* and *east* populations. *S*, *K*, *Q* and *QK* models were used for the *str* population. The quantile-quantile plots indicated that better fits could be observed with different models. For the traits of specific leaf area, branch angle, stem diameter, total height in 2014, total height in 2015 before and after the growing season and Δ^{13}C, the *S* model was the best fit for the *east* population, but the *Q* model was better for the *str* population, and they were selected for further analyses. For nitrogen concentration, the *K* model was the best fit for the *east* population, and the *Q* model for the *str* population. Similarly, for crown width, the *K* model for the *east* population and the *QK* model for the *str* population were selected. For pitch canker resistance, the *K* model for the *total* and *east* populations, and the *K* and *QK* models for the *str* population were selected (Table 3.4). Only the results from the selected models are presented below.

Table 3.4 Selection for the best models for maker-trait associations in *total*, *east* and *str* populations

Traits	***total***		***east***		***str***			
	S[n]	*K*[o]	*S*[n]	*K*[o]	*S*[n]	*Q*[p]	*K*[o]	*QK*[q]
SLA[a]			×			×		
BA[b]			×			×		
CW[c]				×				×
DIA[d]			×			×		
2014H[e]			×			×		
2015HB[f]			×			×		
2015HA[g]			×			×		
Δ^{13}C [h]			×			×		
N[i]				×		×		
PC[j]		×		×			×	×

The best models are marked by ×; [a] Specific leaf area; [b] Branch angle; [c] Crown width; [d] Stem diameter; [e] Total height in 2014; [f] Total height in 2015 before the growing season; [g] Total height in 2015 after the growing season; [h] Carbon isotope discrimination; [i] Nitrogen concentration; [j] Pitch canker resistance; [n] The simple general linear model (GLM); [o] The mixed linear model (MLM) incorporating kinship matrix; [p] The GLM incorporating population structure covariate; [q] The MLM incorporating both kinship matrix and population structure covariate

Associations were identified with specific leaf area (5 SNPs), branch angle (2), crown width (3), stem diameter (4), total height in 2014 (4), total height in 2015 before the growing season (8), total height in 2015 after the growing season (7), Δ^{13}C (4), nitrogen concentration (2) and pitch canker resistance (7) (Table 3.5). When the height-related SNPs were combined, 9 different SNPs were associated with total height. Two loci were detected for both total height and stem diameter, therefore, a total of 34 different SNPs were identified as associated.

The VCFtools software (Danecek et al. 2011) was used to calculate the MAF and to test for Hardy Weinberg Equilibrium (HWE) (Table 3.5). Of the 34 associated SNPs, 4 (12 %) departed from HWE. Fourteen SNPs (41 %) had a MAF greater than 0.05 with a range from 0.07 to 0.32. The MAFs of other SNPs were between 0.01 and 0.04. The effects of these SNPs were relatively small, as r^2 ranged between 0.08 and 0.14. The percentages of clonal and phenotypic variance explained by the SNPs were estimated by comparing the full models (including significant SNPs as random variables in the model) and reduced models (without the significant SNPs in the model). The percentage of clonal and phenotypic variance explained by each locus ranged from 5 % to 27 % and 2 % to 6 %, respectively (Table 3.5). If the individual SNPs for each trait are applied in the models altogether, they can explain 4-18 % of the phenotypic variance (Table 3.7). As the phenotypic data used in this study were collected from only one location, the percentage of each SNP composing the phenotypic variance might be overestimated.

To analyze the additive and dominance effects of the associated SNPs, only the six SNP loci in HWE with all three genotype classes represented were tested using linear regressions (Table 3.6). Additive and dominance effects were significant for all the tested SNP loci at $P < 0.05$, but the estimate for additive effect of locus tscaffold8954_1999_T_G for specific leaf area was not precise as the standard error was greater than half the value of the estimate. Among the six tested loci, the ratio of dominance to additive effects showed the dominance effect was larger in magnitude than the additive effect at four loci. Among them, locus scaffold894573_84188_C_T had the

Table 3.5 SNPs significantly associated with the traits

Trait	SNP[q]	r^2	*p*-value	MAF[j]	PCV[k]	PPV[i]	Location	Candidate gene	Annotation
SLA[a]	scaffold585302_58207_A_G	0.10	2.13E-08	0.02	26.89	2.69	intron	PITA_000065619	Auxin-responsive protein IAA13
SLA[a]	scaffold901438_87380_A_G	0.08	2.24E-07	0.02	22.73	2.28	CDS[r]	PITA_000046166	Predicted, importin subunit alpha-like
SLA[a]	tscaffold7553_225423_C_T	0.09	1.94E-07	0.07	27.27	2.73	CDS[r]	PITA_000013301	Predicted, E3 ubiquitin-protein ligase UPL3
SLA[a]	tscaffold9126_24299_G_A	0.09	9.28E-08	0.03	27.27	2.73	CDS[r]	PITA_000025957	Leucine-rich repeat
SLA[a]	tscaffold8954_1999_T_G	0.09	5.28E-07	0.08	26.14	2.62	P5'RS[n]	PITA_000030967	Putative copper-transporting ATPase HMA5-like isoform X1
BA[b]	scaffold744759_88802_A_G	0.10	4.21E-07	0.08	7.59	2.47	CDS[r]	PITA_000043049	Putative disease resistance protein
BA[b]	scaffold886562_42033_C_T	0.09	1.73E-07	0.02	11.72	3.81	CDS[r]	PITA_000057448	RNA editing factor OTP81
CW[c]	C32326948_8323_G_C	0.12	1.19E-07	0.02	8.05	2.79	UNC[u]	UNC[u]	Kinesin heavy chain[m]
CW[c]	C32326948_24090_G_C	0.12	1.19E-07	0.02	8.05	2.79	3'UTR[t]	PITA_000078936	Putative ATP-dependent RNA helicase
CW[c]	scaffold900647.1_1652_T_C	0.12	1.19E-07	0.01	12.25	4.24	CDS[r]	PITA_000085365	Sec-independent protein translocase protein TatA
DIA[d]	scaffold4572_62040_G_C	0.10	3.67E-07	0.32	13.56	3.72	P5'RS[n]	PITA_000055333	Probable disease resistance protein At4g33300
DIA[d]	scaffold869612_10751_C_T	0.08	2.92E-07	0.04	10.45	2.88	CDS[r]	PITA_000078887	Putative RNA-binding protein
DIA[d]	scaffold894573_84188_C_T	0.09	2.78E-07	0.04	22.57	6.20	CDS[r]	PITA_000057281	Voltage-dependent anion channel
DIA[d]	tscaffold5105_261867_C_T	0.08	5.03E-07	0.02	18.44	5.06	CDS[r]	PITA_000008001	Probable receptor-like serine/threonine-protein kinase At5g57670 isoform X2
TH[e]	C32139602_5870_A_G	0.08	2.60E-07	0.03	11.54	3.45	CDS[r]	PITA_000087565	Predicted, eukaryotic translation initiation factor 5-like
TH[e]	C32495018_13991_T_C	0.11	7.64E-08	0.21	7.69	2.30	P5'RS[n]	PITA_000063143	Glycoside hydrolase family

Table 3.5 Continued

Trait	SNP[q]	r^2	*p*-value	MAF[j]	PCV[k]	PPV[i]	Location	Candidate gene	Annotation
TH[e]	scaffold4572_62040_G_C	0.12	1.81E-08	0.32	12.50	3.70	P5'RS[n]	PITA_000055333	Probable disease resistance protein At4g33300
TH[e]	scaffold802276_312_C_A	0.09	5.15E-07	0.09	18.75	5.56	UNC[u]	UNC[u]	Bifunctional pinoresinol-lariciresinol reductase[m]
TH[e]	scaffold887377_18672_C_T	0.12	8.75E-08	0.26	6.25	1.85	CDS[r]	PITA_000087672	Unknown
TH[e]	scaffold898261_152482_C_T	0.08	2.50E-07	0.02	18.75	5.56	5'UTR[t]	PITA_000028701	Rab family GTPase; Small GTPase superfamily,Ras
TH[e]	tscaffold5847_116012_T_C	0.08	3.96E-07	0.01	15.38	4.60	CDS[r]	PITA_000020895	Predicted, lysosomal beta glucosidase
TH[e]	tscaffold5105_261867_C_T	0.08	2.24E-07	0.02	15.38	4.60	CDS[r]	PITA_000008001	Probable receptor-like serine/threonine-protein kinase At5g57670 isoform X2
TH[e]	tscaffold7910_321796_C_T	0.11	1.77E-07	0.13	7.69	2.30	intron	PITA_000014254	*Copia*-type polyprotein,putative
$\Delta^{13}C$ [f]	scaffold17165_86468_G_A	0.09	1.71E-07	0.01	10.00	6.25	UNC[u]	UNC[u]	Polyadenylate binding protein[m]
$\Delta^{13}C$ [f]	scaffold75093.3_123328_G_A	0.09	1.79E-07	0.01	10.00	6.25	CDS[r]	PITA_000037598	Predicted,histone-lysine N-methyltransferase
$\Delta^{13}C$ [f]	scaffold109938_8086_T_C	0.09	1.78E-07	0.01	5.28	3.13	UNC[u]	UNC[u]	Organic cation/carnitine transporter[m]
$\Delta^{13}C$ [f]	tscaffold4948_129087_G_A	0.10	3.99E-07	0.01	5.00	3.13	CDS[r]	PITAhm_003076	Predicted, small heat shock protein, chloroplastic,
N[g]	C31890840_9044_C_T	0.10	1.76E-07	0.14	9.12	4.20	UNC[u]	UNC[u]	F-box/kelch-repeat protein At3g06240-like[m]
N[g]	C32008620_13926_T_C	0.11	4.70E-07	0.12	8.95	4.12	CDS[r]	PITA_000091703	C3H4 type zinc finger protein
PC[h]	C32571366_88829_G_A	0.14	3.67E-07	0.19	8.82	3.42	CDS[r]	PITA_000019796	Predicted, Leucine-rich repeat
PC[h]	scaffold394991_238149_C_A	0.12	3.98E-07	0.09	7.44	2.89	UNC[u]	UNC[u]	Probable quinone oxidoreductase[m]

Table 3.5 Continued

Trait	SNP[q]	r^2	*p*-value	MAF[j]	PCV[k]	PPV[i]	Location	Candidate gene	Annotation
PC[h]	scaffold490830_154983_T_A	0.14	1.31E-08	0.01	14.84	5.77	CDS[r]	PITA_000030834	Predicted, probable peptide/nitrate transporter At3g53960-like
PC[h]	scaffold771268_7109_G_A	0.13	2.01E-07	0.08	13.14	5.10	P3'RS[o]	PITA_000095986	Predicted, probable aminotransferase TAT2-like
PC[h]	tscaffold758_520984_C_T	0.12	2.98E-07	0.17	7.14	2.85	CDS[r]	PITA_000009116	Unknown
PC[h]	tscaffold1272_36798_C_T	0.11	3.31E-07	0.01	8.47	3.29	intron	PITAhm_002502	Unknown
PC[h]	tscaffold5506_305141_G_A	0.11	2.07E-07	0.01	11.49	4.46	CDS[r]	PITA_000020926	Predicted, dehydration-responsive protein RD22-like isoform 2

[a] Specific leaf area; [b] Branch angle; [c] Crown width; [d] Stem diameter; [e] Total height; [f] Carbon isotope discrimination; [g] Nitrogen concentration; [h] Pitch canker resistance; [i] The percentage of phenotypic variance accounted for by each SNP; [j] Minor allele frequency; [k] The percentage of clonal variance accounted for by each SNP; [m] SNPs that were not located on annotated sequences, instead, the flanking sequences around SNPs were used to query against the NCBI Genbank non-redundant protein database using Blastx; [n] Putative 5' regulatory sequence; [o] Putative 3' regulatory sequence; [q] SNPs were named using scaffold names with the SNP position number in the nucleotide sequence followed by the major and minor SNP alleles; [r] Coding sequences; [t] Untranslated sequences; [u] Unclassified

Table 3.6 Additive and dominance effects for the SNP loci detected by association

SNP[a]	**Trait**	**Additive effect[b]**	**Standard error of the additive effect**	**Dominance effect[b]**	**Standard error of the dominance effect**	**Dominance/ additive**
scaffold894573_84188_C_T	DIA[c]	0.006	0.001	0.027	0.007	4.500
scaffold887377_18672_C_T	2015HA[d]	0.093	0.032	0.11	0.039	1.183
scaffold887377_18672_C_T	2015HB[e]	0.094	0.034	0.114	0.041	1.213
scaffold802276_312_C_A	2014H[f]	-0.185	0.032	-0.364	0.105	1.968
C32008620_13926_T_C	N[g]	0.016	0.006	-0.035	0.02	-2.188
tscaffold8954_1999_T_G	SLA[h]	-0.11	0.105	-0.256	0.108	2.327
C32571366_88829_G_A	PC[i]	0.132	0.031	0.158	0.036	1.197

[a] SNPs were named using scaffold names with the SNP position number in the nucleotide sequence followed by the major and minor SNP alleles; [b] Additive and dominance effects were significantly different from zero based on ANOVA (*P* < 0.05); [c] Stem diameter; [d] Total height in 2015 after the growing season; [e] Total height in 2015 before the growing season; [f] Total height in 2014; [g] Nitrogen concentration; [h] Specific leaf area; [i] Pitch canker resistance

Table 3.7 Percentage of phenotypic variance explained by the associated SNPs and SNP-SNP interactions

Trait	Number of associated SNPs	Phenotypic variance explained by all the associated SNPs, %	Number of epistatic SNP-SNP interactions	Phenotypic variance explained by all the epistatic SNP-SNP interactions, %	Phenotypic variance explained by SNPs from association and epistasis, %
SLA[a]	5	6.03	NA	NA	NA
BA[b]	2	6.43	1	9.08	14.08
CW[c]	3	3.67	2	16.32	17.18
DIA[d]	4	11.73	NA	NA	NA
2014H[e]	4	12.96	2	11.11	14.81
2015HB[f]	7	12.64	NA	NA	NA
2015HA[g]	6	8.55	NA	NA	NA
$\Delta^{13}C$ [h]	4	12.50	2	15.63	21.88
N[i]	2	6.39	1	9.65	13.93
PC[j]	7	17.52	3	13.82	23.36

[a] Specific leaf area; [b] Branch angle; [c] Crown width; [d] Stem diameter; [e] Total height in 2014; [f] Total height in 2015 before the growing season; [g] Total height in 2015 after the growing season; [h] Carbon isotope discrimination; [i] Nitrogen concentration; [j] Pitch canker resistance

largest dominance effect resulting in an increase in stem diameter. At the remaining loci, dominance effects showed a similar magnitude to additive effects.

3.3.4 Epistasis Analyses

Eleven SNP-SNP interactions were identified in associations with branch angle (1 SNP-SNP interaction), crown width (2), total height in 2014 (2), $\Delta^{13}C$ (2), nitrogen concentration (1), and pitch canker resistance (3) (Table 3.8). None of the epistatic loci were detected in the marker-trait association test. All the identified SNPs were in HWE. The loci scaffold901027_91326_G_C and tscaffold6745_463233_G_A were found to be involved in two identified interactions. Sixteen SNPs (80 %) had a MAF greater than 0.05.

Each SNP-SNP interaction accounted for 15-30 % of the clonal variance and 8-10 % of the phenotypic variance for the associated trait (Table 3.8). The combined SNP interaction effects for each trait accounted for 9-16 % of the phenotypic variance (Table 3.7). The addition of these epistatic loci to the associated loci increased the explained proportion of phenotypic variance by 2-14 % (Fig. 3.3). All the identified interactions were between SNPs located on different scaffolds. The most significant SNP-SNP interaction was observed between the scaffold892137_41285_G_A and scaffold461440_154634_T_A loci. Their interaction was related to branch angle. The genotype combination of these two loci showed distinct phenotypic differences (Fig. 3.4).

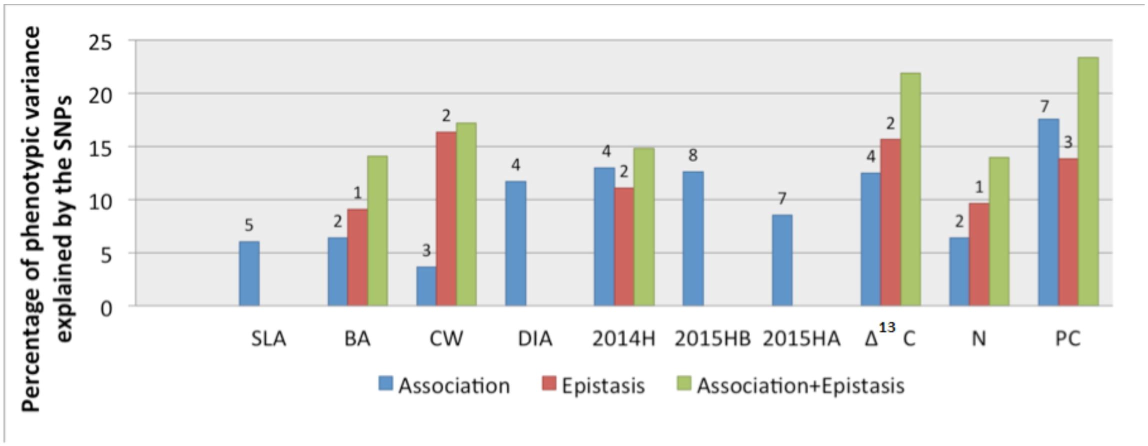

Fig. 3.3 Percentage of phenotypic variance for each trait contributed by the SNPs detected by association and epistasis. The numbers of the identified SNPs (association) or SNP-SNP interactions (epistasis) are presented above the bars. SLA - specific leaf area, BA - branch angle, CW - crown width, DIA - stem diameter, 2014H - total height in 2014, 2015HB and 2015HA - total height in 2015 before and after the growing season, respectively, Δ^{13}C - carbon isotope discrimination, N - nitrogen concentration, and PC - pitch canker resistance

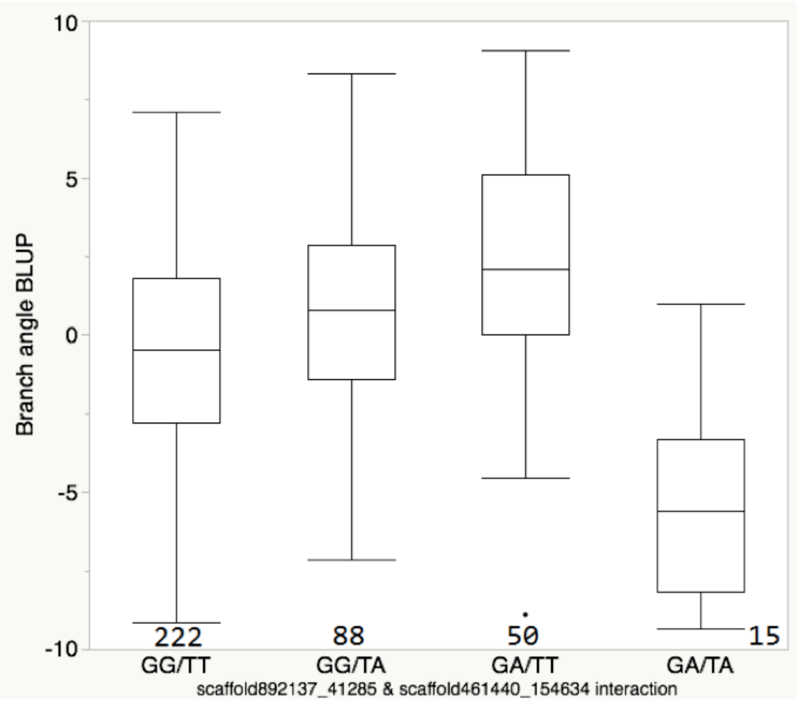

Fig. 3.4 Phenotypic differences between genotype combinations of the loci scaffold892137_41285_G_A and scaffold461440_154634_T_A. The interaction of these two loci is in association with branch angle. Numbers of individuals with the genotype combinations are at the bottom of the bars. The y-axis represents the BLUP estimates of branch angle for the individuals within the population

Table 3.8 SNP-SNP interactions associated with the traits

Trait	SNP1 (MAF)[g]	Location	Candidate gene / Annotation	SNP2 (MAF)[g]	Location	Candidate gene / Annotation	PCV[h]	PPV[i]	*p*-value
BA[a]	scaffold892137_41285_G_A (0.09)	CDS[j]	PITA_000033000 / Predicted, kinesin-like protein KIF18B	scaffold461440_154634_T_A (0.14)	intron	PITA_000021259 / Unknown	27.97	9.09	7.74E-17
CW[b]	scaffold845670_729_G_A (0.09)	UNC[u]	unclassified / Bark storage protein A-like[l]	C32549530_117801_G_T (0.04)	P3'RS[h]	PITA_000046419 / Predicted, wound-induced protein 1-like	28.52	9.87	1.45E-14
CW[b]	scaffold892370.2_38187_G_A (0.16)	CDS[j]	PITA_000068396 / Disease resistance-like protein	scaffold893009.2_42867_G_C (0.19)	CDS[j]	PITA_000068998 / Unknown	23.01	7.97	1.41E-14
2014H[c]	C32560776_5010_G_A (0.09)	CDS[j]	PITA_000038233 / TIR/NBS/LRR disease resistance protein	tscaffold6745_463233_G_A (0.12)	intron	PITA_000009136 / ADP-ribosylation factor	29.53	8.62	9.93E-16
2014H[c]	C32560776_5095_A_G (0.09)	intron	PITA_000038233 / TIR/NBS/LRR disease resistance protein	tscaffold6745_463233_G_A (0.12)	intron	PITA_000009136 / ADP-ribosylation factor	27.21	7.94	6.02E-15
Δ^{13}C [d]	tscaffold488_11868_C_T (0.03)	CDS[j]	PITA_000007619 / Indole-3-acetic acid-amido synthetase GH3.6 (GH3 auxin-responsive promoter)	scaffold894878_57816_T_C (0.48)	intron	PITA_000057030 / Beta-1,3-n-acetylglucosaminyltransferase radical fringe	15.53	9.81	9.22E-15
Δ^{13}C [d]	scaffold117762.1_74566_C_G (0.04)	CDS[j]	PITA_000056667 / Predicted, cytochrome c oxidase subunit 6b-1-like isoform X2	tscaffold2954_107287_G_A (0.10)	intron	PITA_000004221 / Predicted, probable phosphoinositide phosphatase SAC9	16.67	10.53	1.52E-14
N[e]	scaffold90489_7681_G_A (0.08)	UNC[u]	unclassified / Leucine-rich repeat receptor-like serine/threonine-protein kinase BAM2[l]	scaffold807725_59239_A_T (0.28)	CDS[j]	PITA_000043336 / Predicted, uncharacterized zinc finger protein At4g06634 isoform X2	20.95	9.65	3.69E-15

Table 3.8 Continued

Trait	SNP1 (MAF)[g]	Location	Candidate gene / Annotation	SNP2 (MAF)[g]	Location	Candidate gene / Annotation	PCV[h]	PPV[i]	*p*-value
PC[f]	C31644976_3483_C_T (0.03)	UNC[u]	unclassified / Putative leucine-rich repeat receptor-like serine/threonine-protein kinase[l]	scaffold898450_46698_G_T (0.21)	intron	PITA_000069402 / Pentatricopeptide repeat-containing protein At1g12300, mitochondrial-like	22.81	8.86	2.69E-15
PC[f]	scaffold901027_91326_G_C (0.05)	CDS[j]	PITA_000045299 / Transcription factor	tscaffold2366_165430_C_T (0.14)	3'UTR[i]	PITA_000012460 / Unknown	23.19	9.01	7.75E-15
PC[f]	scaffold901027_91326_G_C (0.05)	CDS[j]	PITA_000045299 / Transcription factor	tscaffold362_1145764_A_T (0.14)	Intron	PITAhm_003320 / Mitochondrial processing peptidase subunit beta; GBFi nteracting protein 1-like	22.60	8.78	4.57E-15

[a] Branch angle; [b] Crown width; [c] Total height in 2014; [d] Carbon isotope discrimination; [e] Nitrogen concentration; [f] Pitch canker resistance; [g] SNPs were named using scaffold names with the SNP position number in the nucleotide sequence followed by the major and minor SNP alleles; [h] The percentage of clonal variance accounted for by each SNP; [i] The percentage of phenotypic variance accounted for by each SNP; [j] Coding sequences; [h] Putative 3' regulatory sequence; [i] Untranslated sequences; [l] SNPs that were not located on annotated sequences, instead, the flanking sequences around SNPs were used to query against the NCBI Genbank non-redundant protein database using Blastx; [u] Unclassified

3.3.5 Annotation of Genes that Contained SNPs Associated with Traits

The loci with the identified SNPs were annotated using the loblolly pine Gene Annotation 3.0 (http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/Pita_Annotation_v3.0/) (Wegrzyn et al. 2014). Of the 34 SNPs identified by association, 19 resided in CDS, 2 in UTRs, 3 in introns, 4 in putative regulatory sequences and 6 in unannotated (unclassified) regions (Table 3.5). Of the 20 SNPs identified in the epistasis analysis, 8 resided in CDS, 7 in introns, 1 in a UTR, 1 in a putative regulatory sequence, and 3 in unannotated regions (Table 3.8).

For the identified SNPs located in tentative genes, the function annotation was done and presented in Tables 3.5 and 3.8. Otherwise, the flanking sequences around the SNPs were used for a Blastx search to identify the functions of genes that contained SNPs associated with traits. Associations between crown structure traits and genes related to developmental regulation, transport and stress response were detected. Genes involved in RNA editing (RNA editing factor OTP81) and disease resistance (putative disease resistance protein) were found to be associated with branch angle. An interaction between genes encoding a microtubule dynamics regulation protein (kinesin-like protein KIF18B) (Tanenbaum et al. 2011) and an unknown protein may also affect branch angle. Genes involved in auxin-mediated development (auxin-responsive protein IAA13), pathogen-recognizing disease resistance and cell wall developmental processes (leucine-rich repeat) (Draeger et al. 2015), nuclear protein import mediation (importin subunit alpha-like)(Hübner et al. 1999), ubiquitin transfer (E3 ubiquitin-protein ligase UPL3)

(Berndsen and Wolberger 2014) and copper transport (copper-transporting ATPase HMA5), were associated with specific leaf area. Genes encoding products involved in cell division (kinesin heavy chain) (Yang et al. 1989), unwinding of the RNA helix (putative ATP-dependent RNA helicase) and protein transport (sec-independent protein translocase protein TATA) were associated with crown width. An interaction between genes encoding a bark storage protein (bark storage protein A-like) and a wound-induced protein, and an interaction between genes encoding a disease resistance protein and an unknown protein were detected to be associated with crown width.

One interesting association with total height was found for the locus scaffold802276_312_C_A. The SNP was found in a gene encoding a bifunctional pinoresinol-lariciresinol reductase, which is involved in lignan biosynthesis (Renouard et al. 2014). Additionally, a broad spectrum of genes, encoding regulation and defense functions were also associated with height, including the genes related to vesicle trafficking (Rab family GTPase; Small GTPase superfamily, Ras), disease resistance (probable receptor-like serine/threonine-protein kinase At5g57670 isoform X2; probable disease resistance protein At4g33300), biosynthesis and degradation of glycogen (glycoside hydrolase family), retrotransposition (putative copia-type polyprotein), and translation initiation (probable eukaryotic translation initiation factor 5-1). An interaction between a plant resistance R-gene (putative TIR/NBS/LRR disease resistance protein) and a regulator gene of vesicular traffic and actin remodeling (ADP-ribosylation factor) (Pasqualato et al. 2002) was also found to affect height.

Some of the genes associated with total height were also associated with stem diameter, including genes involved in disease resistance (probable disease resistance protein At4g33300; probable receptor-like serine/threonine-protein kinase At5g57670 isoform X2). In addition, genes encoding voltage-dependent anion channels and RNA-binding protein were also associated with stem diameter.

For $\Delta^{13}C$, one interesting associated SNP was tscaffold4948_129087_G_A, which located in a gene encoding a small heat shock protein (sHSP), suggesting that a stress response mechanism might participate in the regulation of WUE. Genes encoding proteins involved in RNA–protein complexes (polyadenylate binding protein), epigenetic regulation (histone-lysine N-methyltransferase) and transportation of organic cations and carnitine (organic cation/carnitine transporter) were also associated with $\Delta^{13}C$. An interaction between genes encoding an IAA-amido synthase that conjugates aspartic acid and other amino acids to auxin (indole-3-acetic acid-amido synthetase GH3.6) and a transferase (beta-1,3-n-acetylglucosaminyltransferase radical fringe) was found to also contribute to $\Delta^{13}C$. In addition, interaction between genes involved in electron transport chain (cytochrome c oxidase subunit 6b-1-like isoform X2) and degradation of phosphoinositide signals (phosphoinositide phosphatase SAC9) was also associated with $\Delta^{13}C$.

Genes encoding a F-box/kelch-repeat protein and a zinc finger protein were found to associate with nitrogen concentration. A plant meristem and organ developmental gene (leucine-rich repeat receptor-like serine/threonine-protein kinase BAM2 or LRR RLK BAM2) and a zinc finger protein gene were found to interact.

For pitch canker resistance, transporter (probable peptide/nitrate transporter At3g53960-like), metabolism (probable quinone oxidoreductase; probable aminotransferase TAT2-like), and stress-induced genes (leucine-rich repeat; dehydration-responsive protein RD22-like isoform 2) were found to be associated. Genes related to disease resistance (putative leucine-rich repeat receptor-like serine/threonine-protein kinase), transcription factor, organelle gene expression regulation (pentatricopeptide repeat-containing protein At1g12300, mitochondrial-like; mitochondrial processing peptidase subunit beta) were found in the interactions associated with pitch canker resistance.

We investigated the effect of SNP substitution on amino acid sequence by aligning and comparing the amino acid sequences translated from the sequences with the different alleles and the corresponding transcripts (Table 3.9). Of the 27 SNPs resided in CDS, 19 (70 %) caused nonsynonymous substitutions. Two nonsynonymous substitutions generated stop codon and caused premature truncation of the coding sequences, while the others resulted in amino acid replacements. One nonsynonymous substitution resulting in premature truncation occurs on the SNP locus associated with branch angle. It is located in a gene that encodes a RNA editing factor. Another one occurs on the locus associated with stem diameter and total height. It is located in a gene encoding a receptor-like serine/threonine-protein kinase.

Table 3.9 Nonsynonymous and synonymous SNP substitutions

Trait	**SNP**[k]	**Substitution type**
Detected from association		
SLA[a]	scaffold901438_87380_A_G	I to V
SLA[a]	tscaffold7553_225423_C_T	S to L
SLA[a]	tscaffold9126_24299_G_A	E to K
BA[b]	scaffold886562_42033_C_T	Q to stop
BA[b]	scaffold744759_88802_A_G	synonymous
CW[d]	scaffold900647.1_1652_T_C	P to L
DIA[e]	scaffold869612_10751_C_T	S to L
DIA[e]	scaffold894573_84188_C_T	T to I
DIA/TH	tscaffold5105_261867_C_T	R to stop
TH[f]	scaffold887377_18672_C_T	A to V
TH[f]	C32139602_5870_A_G	K to R
TH[f]	tscaffold5847_116012_T_C	W to R
$\Delta^{13}C$ [g]	scaffold75093.3_123328_G_A	synonymous
$\Delta^{13}C$ [g]	tscaffold4948_129087_G_A	synonymous
N[h]	C32008620_13926_T_C	synonymous
PC[i]	C32571366_88829_G_A	synonymous
PC[i]	scaffold490830_154983_T_A	I to K
PC[i]	tscaffold758_520984_C_T	A to V
PC[i]	tscaffold5506_305141_G_A	synonymous
Detected from epistasis		
BA[b]	scaffold892137_41285_G_A	synonymous
CW[d]	scaffold892370.2_38187_G_A	synonymous
CW[d]	scaffold893009.2_42867_G_C	A to P
2014H[j]	C32560776_5010_G_A	D to G
$\Delta^{13}C$ [g]	tscaffold488_11868_C_T	V to A
$\Delta^{13}C$ [g]	scaffold117762.1_74566_C_G	L to V

Table 3.9 Continued

Trait	**SNP**[k]	**Substitution type**
N[h]	scaffold807725_59239_A_T	N to Y
PC[i]	scaffold901027_91326_G_C	R to T

[a] Specific leaf area; [b] Branch angle; [d] crown width; [e] Stem diameter; [f] Total height; [g] Carbon isotope discrimination; [h] Nitrogen concentration; [i] Pitch canker resistance; [j] Total height in 2014; [k] SNPs were named using scaffold names with the SNP position number in the nucleotide sequence followed by the major and minor SNP allele

3.4 Discussion

3.4.1 Broad Genetic Correlations

The clonal repeatability was high for all the measured traits except for height growth. However, with the exception of total height, the traits in this study were only measured on one population in one year and in one location, and therefore, the clonal repeatability estimates were up-biased. Nonetheless, previous studies have shown these traits to be heritable in loblolly pine. Emhart et al. (2007) determined the broad-sense heritabilities of crown radius, leaf area and branch angle in loblolly pine to be 0.20, 0.25 and 0.26, respectively, when estimated from a combined analysis across families. Baltunis et al. (2008) reported that the across-site estimate of broad-sense heritability for Δ^{13}C was 0.19 and Emhart (2005) reported similar estimates 0.23 and 0.17 based on two separate years of sampling. For nitrogen concentration, Cumbie et al. (2011) reported that the broad-sense clone mean heritability was 0.42. Pitch canker resistance in loblolly pine is also a heritable and complex trait with a continuous distribution across clones (Kayihan et al. 2005). Quesada et al. (2010) estimated 30-40 % of the disease variation could be attributed to genetic effects. The clonal repeatability and the considerable phenotypic variation suggested the population used in this study was suitable for association mapping. The lower clonal repeatability for height growth during the 2015 growing season agrees with previous conclusions that growth rate has low heritability (White et al. 2007; Shmulsky and Jones 2011) and is more affected by environmental effects, such as availability of light, water and nutrition rather than the genetic components. Therefore, height growth in 2015 was excluded from further analyses.

Strong positive correlations between crown width, height and stem diameter and a weak negative correlation between crown width and branch angle indicated that bigger trees tended to have wider crowns and flatter branches, which is in agreement with previous progeny tests of loblolly pine measuring nine or ten year old trees on four sites (Lambeth and Hubert 1997). The wider crown and flatter branches enabled the trees to capture light better and to be more competitive than other trees, thus accumulating more biomass. The genetic correlations of crown width with other growth traits along with its medium to high heritability as reported in the previous studies suggested crown width could be a key component of productivity, and selection for crown width could favor growth traits (Lambeth and Hubert 1997; Emhart et al. 2007). Genes affecting crown structure in loblolly pine have been rarely explored in the past. Considering heritability of crown structure and growth traits, molecular markers associated with these traits could be valuable in marker assisted selection (MAS).

Due to a changing climate, forest trees with better adaptive characteristics such as superior photosynthetic and water use abilities will be needed in the future (IPCC 2014a). In this study, slight positive correlations existed between nitrogen concentration and specific leaf area, crown width and total height, suggesting a higher nitrogen concentration may have increased the leaf area and tree size through a promoted photosynthetic ability since nitrogen is indispensable for Rubisco, a key enzyme of the Calvin cycle (Bloomfield et al. 2014). It should be noted that the correlations shown here for nitrogen concentration as well as Δ^{13}C were subtle, possibly because the trees contained in this population were originally from a broad geographic range, and these different genotypes may display distinct photosynthetic and water use strategies depending upon environments (Flanagan and Johnsen 1995).

Stable Δ^{13}C in plants reflects the balance between photosynthetic ability and stomatal conductance. It has long been used as a measure for WUE in forest trees as less discrimination is associated with higher WUE (Aitken et al. 1995; Baltunis et al. 2008). The trees with flatter branch angles, wider crowns, greater heights and higher nitrogen concentrations tended to have lower Δ^{13}C, suggesting the fast-growing trees with higher light capture and photosynthetic ability have a better WUE. It is worth noting that these trees may also exhibit greater stomatal closure under high vapour pressure deficit (VPD) or limiting soil moisture since they are deploying large canopies and have greater nitrogen concentration. These canopies are perhaps more productive over some time period, but would be (perhaps) more vulnerable to hydraulic failure.

The empirical relationship between Δ^{13}C and growth can be negative, positive or uncorrelated depending on the specific environment or species (Orians and Solbrig 1977; Aitken et al. 1995; Flanagan and Johnsen 1995; Li et al. 2013). It can be hypothetically explained by a hypothesis that high WUE could be at the expense of growth (Orians and

Solbrig 1977). However, neither our study, which was conducted in seasons with normal precipitation, nor the study conducted by Cumbie et al. (2011) during seasons with drought supported this hypothesis for loblolly pine. It should be noted that in both studies, relatively young samples were measured. Our trees were measured in their fourth growing season, and those in the Cumbie et al. study in their second growing season. It is possible that young trees may use a different WUE strategy (Aitken et al. 1995).

3.4.2 Geographical Variation

Environmental heterogeneity and gradients drive the adaptation of forest trees, thus creating geographical variation within the natural range. Through adaptive and selectively neutral processes, loblolly pine developed geographic differences between populations east and west of the Mississippi River due to the 100,000-year refugia isolation. Reports have shown that loblolly pines from west of the Mississippi River are slower growing but more resistant to aridity and crowding (Schmidtling 2001). Our study demonstrated that genotypes with eastern origins tended to grow faster and have a better WUE than western and far west genotypes. The growth rate difference was consistent with previous results (Schmidtling 2001), but the drought tolerance within this population is difficult to judge since relatively young trees were measured and were under normal precipitation conditions during the growing seasons of sampling. Additionally, the studied trees in our population were grown from open pollinated seeds, and only the maternal origins can be determined, with the paternal origins uncertain.

Therefore, the measured trees may have different drought tolerance phenotypes from the originally selected maternal parents.

3.4.3 Non-Additive Effects

Most robustly associated SNPs detected in loblolly pine association studies account for only a fraction of the total genetic variance in a trait (González-Martínez et al. 2007; Quesada et al. 2010). Similarly, in our study, only 5-27 % of the clonal variance and 2-6 % of the phenotypic variance could be explained by the associated SNPs. This agrees with the hypothesis that most quantitative traits are affected by many genes with small effects (Flint and Mackay 2009) as well as previous evidence in loblolly pine (Emhart et al. 2007; Quesada et al. 2010; Cumbie et al. 2011). It is also possible that genes with major effects remain undetected. However, in this study, we analyzed 2,822,609 SNPs in or near 48,391 high quality tentative genes, so it seems unlikely that we would have missed the genes with major effects for every trait if they exist. Dissecting these quantitative traits and revealing their genetic control remain challenging. To better examine these traits, we extended our investigation beyond additive to non-additive effects, namely, dominance and epistasis. Though only a few SNP-SNP interactions were determined to be associated, they generally contributed more to the clonal and phenotypic variance than the additive loci (Fig. 3.3). None of the epistatic loci were discovered in association studies, indicating the additive and epistasis effects may determine the traits using independent networks (Zhang et al. 2015). The dominance effects are similar or larger in magnitude than additive effects. Although the additive effects are the main focus for loblolly pine breeding, the results from this research as well as previous studies indicated that both the dominance and epistasis effects should be considered in MAS of loblolly pine, since they might play important roles in capturing desirable traits (Eckert et al. 2009a; Cumbie et al. 2011). Additionally, only SNPs were investigated in this study, however, other kinds of polymorphisms such as indels, copy number variations, transposable elements and stable epigenetic modifications could also explain the phenotypic variation.

3.4.4 Non-Coding and Rare Variants

The non-coding and rare variants were also used to address quantitative trait dissection problems. In this study, we used exome sequencing to identify over 2.8 million SNPs when filtering conditions were relaxed to include SNPs with a MAF greater than 0.01, assuring a broad spectrum of SNPs containing the rare alleles to be investigated. Moreover, in exome sequencing, the capture often extends to non-target regions, and therefore, variants adjacent to CDS and UTRs, including introns and putative regulatory elements were also identified. Among the loci detected by the association analyses, more than half had a MAF smaller than 0.05 and 44 % resided in non-CDS (Table 3.5). These low frequency and non-CDS variants are important resources for exploring the quantitative traits of loblolly pine. As the reference assembly and gene annotation for the loblolly pine genome are under active improvement, the variant locations and annotations might be revised in the future.

3.4.5 Application of the Identified Variants

MAS utilizing the identified alleles and those that will be discovered in the future using the SNPs identified in this project, may facilitate loblolly pine breeding. For instance, within this population, tree 276B contained the most desired alleles at 18 additive and 2 epistatic loci, which were associated with the traits of total height, stem diameter, Δ^{13}C, nitrogen concentration and pitch canker resistance. This corresponded to its superior performance indicated by its above median values of measurement on growth traits. The height-related loci, which were detected multiple times using measurements at different times, and two alleles detected in association with both stem diameter and total height can be regarded with higher validity than the other loci. However, before these loci can be applied for breeding, they need to be verified using more samples at different ages and with replications in different locations.

3.4.6 Heterozygosity-Trait Correlations

It is hypothesized that individual heterozygosity may correlate with individual fitness and superior trait performance due to dominance or overdominance (heterosis) (Charlesworth and Willis 2009; Ruiz-Lopez et al. 2012; Rodríguez-Quilón et al. 2015). Correlating individual heterozygosity measured using genetic markers with individual fitness related or adaptation-associated traits (heterozygosity-trait correlations or HTCs) could test this. Forest trees have been used to test for HTCs. Ledig et al. (1983) reported a significant correlation between mean annual basal area increment and heterozygosity in pitch pine suggesting that the growth may be positively associated with isozyme heterozygosity. Using a maritime pine (*Pinus pinaster* Ait.) population, no significant correlation was found between survival and genome-wide heterozygosity (Rodríguez-Quilón et al. 2015), nonetheless, the authors suggested the heterozygosity of specific candidate genes was of great importance to increase fitness. In another study focusing on a Siberian larch (*Larix sibirica* Ledeb.) population, no relationship was found between individual heterozygosity and radial growth (Babushkina et al. 2016), but the authors pointed that relationships could be rather complex depending on the tree age, and more markers and samples are needed to address it. In this study, individual heterozygosity was found to be associated with $\Delta^{13}C$ and nitrogen concentration within a clonally tested loblolly pine population. Different correlation results detected using different populations may be due to population size and as the size gets smaller, significance goes down. To verify the effects of heterozygosity on the traits, a population including more samples with all three genotypes present should be tested for the correlation. Additionally, it would be interesting to calculate and compare the HTCs using individual heterozygosity based on supposedly neutral markers and loci under selection in the loblolly pine genome separately.

3.4.7 Putative Functions of Genes that Contained SNPs Associated with Traits

In this study, exome-derived probes were used for sequence capture, hence the SNPs were identified mostly in the gene spaces or very close to them. Since linkage disequilibrium decays rapidly within this population (Lu et al. 2016), the identified SNPs are likely to be within or close to the genes controlling the phenotypic traits. Previous studies have used nearly 4000 SNPs to identify loci associated with height, $\Delta^{13}C$, nitrogen concentration, pitch canker disease resistance and other important traits (Quesada et al. 2010; Cumbie et al. 2011). To validate the SNPs identified in this study, we mapped those sequences with previously identified SNPs to loblolly pine reference assembly v1.01 (http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01) using the software GMAP (Wu and Watanabe 2005). However, none of them could be mapped to the genes or flanking sequences that contained SNPs associated with traits reported in this study. Nonetheless, genes reported in this study provide valuable clues to understand the genetic architecture of complex traits.

Discovery of an associated RNA editing factor gene implies that RNA editing influences branch angle, possibly through providing RNA edited proteins to incorporate into polypeptide complexes (Brennicke et al. 1999). The association of a bark storage protein gene suggests nitrogen resorption is involved in the development of crown width (Zhu and Coleman 2001).

Genes encoding an auxin-responsive protein and a copper-transporting ATPase were found to be associated with specific leaf area. Discoveries of these genes suggest the auxin-regulation participates in the process of leaf meristem growth and determine its size and shape (Zgurski et al. 2005); copper is an essential element for leaf growth since copper deficiency defects in plants include a general reduced growth rate, chlorosis, especially in young leaves, curling of leaf margins (Puig 2014).

One interesting gene that contained SNPs associated with total height encodes bifunctional pinoresinol-lariciresinol reductase. This gene is involved in lignan biosynthesis (Renouard et al. 2014). In trees, lignans are synthesized and deposited in significant amounts in the heartwood region, probably preventing heart rot caused by fungi (Suzuki and Umezawa 2007).

Two interesting genes that contained SNPs associated with $\Delta^{13}C$ encode histone-lysine N-methyltransferase and a small heat shock protein (sHSP), respectively. $\Delta^{13}C$ reflects the water use efficiency of plants, which is regulated by stomatal responses to changes in VPD. Phytohormone abscisic acid (ABA) was found to be the means by which angiosperm stomata respond to natural changes in VPD (McAdam et al. 2016). As reported by Zheng et al. (2012), a histone methyltransferase expression is regulated by ABA; also, Sun et al. (2016) reported a sHSPs may function as a protein chaperone to modulates ABA biosynthesis and ABA signaling. It is possible that these two genes are on the pathway of ABA biosynthesis and signaling, hence impacting the stomatal responses and water use in loblolly pines.

One gene that contained SNPs associated with nitrogen concentration encodes a F-box/kelch-repeat protein. An Arabidopsis F-box/kelch protein is involved in timing of flowering and phenylpropanoid biosynthesis (Zhang et al. 2013). Its association with nitrogen concentration suggests flowering and phenylpropanoid biosynthesis may be under the control of nitrogen concentration. The LRR RLK BAM2 gene was also found to be associated with nitrogen concentration. BAM2 has been recognized for its role in the development of vascular strands within leaves and a correlated control of leaf shape, size and symmetry. A similar receptor-like protein kinase-like protein (RLK) was reported by Cumbie et al. (2011) as a candidate gene for regulation of nitrogen concentration.

For pitch canker resistance, two SNPs located on resistance related genes were identified, one is on a leucine-rich repeat gene, and the other is on a dehydration-responsive protein gene. The leucine-rich repeat domain and the nucleotide-binding site domain are the major parts that compose the main class of R-genes (Leister and Katagiri 2000). When encountered by disease, plants produce R proteins to detect the presence of pathogen effectors, resulting in activation of multiple signaling pathways and transcription of specific genes that limit pathogen proliferation and disease symptom expression (Arango-Velez et al. 2014). A dehydration-responsive gene often acts to suppress the drought stress response (Yamaguchi-Shinozaki and Shinozaki 1993). One syndrome of pitch canker disease infection is the wilt of tips of girdled branches due to obstructed water flow. The association of a dehydration-responsive protein RD22-like gene suggests that water deficit resistance has also been induced as part of the defense mechanism against pitch canker disease. As pointed out before, pitch canker resistance is a quantitative trait influenced by many genes with relatively small effects (Quesada et al. 2010). In this study, the identified resistance related SNPs explain a small proportion of clonal and phenotypic variation, but SNP-SNP interaction analyses explain a higher proportion. A transcription factor gene was identified to interact with a gene encoding a mitochondrial processing peptidase subunit beta. Another interaction was also found to be related to mitochondria. It’s between a gene encoding a leucine-rich repeat receptor-like serine/threonine-protein kinase and a gene encoding a mitochondrial-like pentatricopeptide repeat-containing protein. The latter gene is one of the major mediators for mitochondrial post-transcriptional regulation (Manna 2015). Both above-mentioned interactions imply regulation of mitochondrial gene expression is involved in pitch canker resistance. However, it remains unclear how the mitochondrial gene products perform their functions in the resistance pathway. It does offer new ideas to explore genes related to pitch canker resistance.

4. CONCLUSIONS

Loblolly pine's characteristics such as amenability to plantation management, high yields and fast growth make it one of the most economically important forest species in the world. Timber and pulpwood are the primary products. Loblolly pine is also a promising tool in efforts to relieve warming and long-lasting climate changes caused by greenhouse gas emissions. However, the large and complex genome of loblolly pine poses challenges for tree improvement through gene discovery, association studies, and genomic selection. Genotyping-by-sequencing facilitates the process of genetic variation discovery and the collection of genome-wide molecular markers in an efficient and budget-wise manner. The availability of molecular markers can assist with the characterization of linkage disequilibrium and genomic structure of the population. Additionally, molecular markers including rare and non-coding variants provide great opportunity to dissect complex traits using association mapping and identify the genes and their effects that underlie complex traits.

In Chapter II, exome-sequencing was conducted to discover genetic variation in the clonally tested ADEPT2 loblolly pine population that included clones of 375 loblolly pine trees originally sampled across a wide range. Sequence capture oligonucleotide probes were designed using 199,723 exons in 48,391 high quality tentative genes listed in gene annotation v. 2.0 for loblolly pine genome assembly v. 1.01. The Illumina HiSeq 2500 platform was used to sequence the captured and enriched libraries for each tree. Nearly 99 % of the sequence reads were mapped to the reference genome assembly. The capture efficiency and specificity were high as 67 % of the reads per tree mapped to the capture target regions. Among all the trees, at least 83 % of the capture target bases had coverage of 5X, 72 % - 10X, and 49 % - 20X. With the filtering condition of being bi-allelic sites with at least 10X sequencing depth in at least 90% of the individuals and with the MAF ≥ 0.05, a total of 972,720 SNPs were acquired for downstream analyses. Analyses of heterozygosity and F_{IS} indicated this population is highly heterozygous with a low inbreeding rate and a high level of genetic diversity. The average LD for linked SNPs was inferred from the trendlines of the nonlinear regressions and started from 0.44, then decayed by half (0.22) at 55 bp, and to 0.10 at 192 bp. LD decayed faster than previously reported suggesting that a great number of markers will be required for association mapping. Genomic structure analyses showed this population consists of two distinct subpopulations (genetic clusters), west and east of the Mississippi River.

In Chapter III, over 2.8 million SNP markers identified and genotyped by exome capture and sequencing were used to test for correlations of individual heterozygosity, single locus associations, and SNP-SNP interactions with phenotypic traits. Within the tested loblolly pine population, numerous genetic correlations between traits were detected as well as geographical variation. Individual heterozygosity was found to potentially correlate with $\Delta^{13}C$ and nitrogen concentration. Thirty-four SNPs and eleven SNP interactions were associated with crown structure, growth, physiology and disease resistance traits among more than 2.8 million SNPs identified and genotyped by exome capture and sequencing. Dominance and epistatic effects were substantial complements to additive effects. These results provide direction for loblolly pine breeding strategy improvement through MAS and genomic selection. Candidate genes with a broad spectrum of functions were identified.

Our results demonstrated the efficiency of exome capture for genotyping a species with a large, complex genome. The highly diverse genetic variation reported in this study provides a valuable resource for loblolly pine breeding through MAS and genomic selection. Association studies and the functional analyses of the promising candidate genes will facilitate elucidation of the genetic architecture of the loblolly pine traits and contribute molecular tools for selection of loblolly pine genotypes adapted to changing climate scenarios.

REFERENCES

Adams WT, Joly RJ (1980) Linkage relationships among twelve allozyme loci in loblolly pine. J Hered 71:199-202

Aitken SN, Kavanagh KL, Yoder BJ (1995) Genetic variation in seedling water-use efficiency as estimated by carbon isotope ratios and its relationship to sapling growth in Douglas-fir. Forest Genetics 2:199-206

Al-Rabab'ah M, Williams C (2002) Population dynamics of *Pinus taeda* L. based on nuclear microsatellites. Forest Ecol Manag 163:263-271

Arango-Velez A, González LMG, Meents MJ, El Kayal W, Cooke BJ, Linsky J, Lusebrink I, Cooke JE (2014) Influence of water deficit on the molecular responses of *Pinus contorta*× *Pinus banksiana* mature trees to infection by the mountain pine beetle fungal associate, *Grosmannia clavigera*. Tree Physiol 34:1220-1239. doi: 10.1093/treephys/tpt101

Babushkina EA, Vaganov EA, Grachev AM, Oreshkova NV, Belokopytova LV, Kostyakova TV, Krutovsky KV (2016) The effect of individual genetic heterozygosity on general homeostasis, heterosis and resilience in Siberian larch (*Larix sibirica* Ledeb.) using dendrochronology and microsatellite loci genotyping. Dendrochronologia 38:26-37

Baker JB, Langdon OG (1990) *Pinus taeda* L., loblolly pine In: Burns RM, Honkala BH (eds) Silvics of North America, vol 1, Confiers. Agriculture handbook no.654. U.S. Department of Agriculture, Forest Service, Washington, DC, pp 497-512

Baltunis B, Martin T, Huber D, Davis J (2008) Inheritance of foliar stable carbon isotope discrimination and third-year height in *Pinus taeda* clones on contrasting sites in Florida and Georgia. Tree Genet Genomes 4:797-807. doi:10.1007/s11295-008-0152-2

Bellassen V, Luyssaert S (2014) Carbon sequestration: managing forests in uncertain times. Nature 506:153-155

Benomar L, Lamhamedi MS, Rainville A, Beaulieu J, Bousquet J, Margolis HA (2016) Genetic adaptation vs. ecophysiological plasticity of photosynthetic-related traits in young *Picea glauca* trees along a regional climatic gradient. Front Plant Sci 7:48. doi:10.3389/fpls.2016.00048

Berndsen CE, Wolberger C (2014) New insights into ubiquitin E3 ligase mechanism. Nat Struct Mol Biol 21:301-307. doi:10.1038/nsmb.2780

Bloomfield KJ, Farquhar GD, Lloyd J (2014) Photosynthesis-nitrogen relationships in tropical forest tree species as affected by soil phosphorus availability: a controlled environment study. Funct Plant Biol 41:820-832. doi:10.1071/Fp13278

Bolte A, Ammer C, Lof M, Madsen P, Nabuurs GJ, Schall P, Spathelf P, Rock J (2009) Adaptive forest management in central Europe: Climate change impacts, strategies and integrative concept. Scand J Forest Res 24:473-482. doi:10.1080/02827580903418224

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635. doi:10.1093/bioinformatics/btm308

Brautigam K, Vining KJ, Lafon-Placette C, Fossdal CG, Mirouze M, Marcos JG, Fluch S, Fraga MF, Guevara MA, Abarca D, Johnsen O, Maury S, Strauss SH, Campbell MM, Rohde A, Diaz-Sala C, Cervera MT (2013) Epigenetic regulation of adaptive responses of forest tree species to the environment. Ecol Evol 3:399-415. doi:10.1002/ece3.461

Brennicke A, Marchfelder A, Binder S (1999) RNA editing. FEMS Microbiol Rev 23:297-316

Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U S A 101:15255-15260. doi:10.1073/pnas.0404231101

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097. doi:10.1086/521987

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7. doi:10.1186/s13742-015-0047-8

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nat Rev Genet 10:783-796. doi:10.1038/nrg2664

Chhatre VE, Byram TD, Neale DB, Wegrzyn JL, Krutovsky KV (2013) Genetic structure and association mapping of adaptive and selective traits in the east Texas loblolly pine (*Pinus taeda* L.) breeding populations. Tree Genet Genomes 9:1161-1178. doi:10.1007/s11295-013-0624-x

Conkle MT (1981) Isozyme variation and linkage in six conifer species. In: Gen.Tech.Rep.PSW-GTR-48. Pacific Southwest Forest and Range Exp. Stn, Forest Service, U.S. Department of Agriculture, Berkeley,CA, pp 11-17

Cortijo S, Wardenaar R, Colome-Tatche M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Hospital F, Aury JM, Wincker P, Roudier F, Jansen RC, Colot V, Johannes F (2014) Mapping the epigenetic basis of complex traits. Science 343:1145-1148. doi:10.1126/science.1248127

Cregg B, Zhang J Carbon isotope discrimination as a tool to screen for improved drought tolerance. In: 11th Metropolitan Tree Improvement Alliance (METRIA) Conference, Gresham, Oregon, 2000.

Cumbie WP, Eckert A, Wegrzyn J, Whetten R, Neale D, Goldfarb B (2011) Association genetics of carbon isotope discrimination, height and foliar nitrogen in a natural population of *Pinus taeda* L. Heredity (Edinb) 107:105-114. doi:10.1038/hdy.2010.168

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Analysis G (2011) The variant call format and VCFtools. Bioinformatics 27:2156-2158. doi:10.1093/bioinformatics/btr330

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592-601. doi:10.1016/j.tplants.2014.05.006

Devey ME, Fiddler TA, Liu BH, Knapp SJ, Neale DB (1994) An RFLP linkage map for loblolly pine based on a three-generation outbred pedigree. Theor Appl Genet 88:273-278. doi:10.1007/BF00223631

Draeger C, Ndinyanka Fabrice T, Gineau E, Mouille G, Kuhn BM, Moller I, Abdou MT, Frey B, Pauly M, Bacic A, Ringli C (2015) Arabidopsis leucine-rich repeat extensin (LRX) proteins modify cell wall composition and influence plant growth. BMC Plant Biol 15:155. doi:10.1186/s12870-015-0548-8

Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, St Clair JB, Neale DB (2009a) Association genetics of coastal Douglas fir (*Pseudotsuga menziesii* var. *menziesii*, Pinaceae). I. Cold-hardiness related traits. Genetics 182:1289-1302. doi:10.1534/genetics.109.102350

Eckert AJ, Pande B, Ersoz ES, Wright MH, Rashbrook VK, Nicolet CM, Neale DB (2009b) High-throughput genotyping and mapping of single nucleotide polymorphisms in loblolly pine (*Pinus taeda* L.). Tree Genet Genomes 5:225-234. doi:10.1007/s11295-008-0183-8

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, Gonzalez-Martinez SC, Neale DB (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185:969-982. doi:10.1534/genetics.110.115543

Eckert AJ, Wegrzyn JL, Cumbie WP, Goldfarb B, Huber DA, Tolstikov V, Fiehn O, Neale DB (2012) Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. New Phytol 193:890-902. doi:10.1111/j.1469-8137.2011.03976.x

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379. doi:10.1371/journal.pone.0019379

Emhart VI (2005) Physiological genetics of contrasting loblolly and slash pine families and clones. Dissertation, University of Florida

Emhart VI, Martin TA, White TL, Huber DA (2007) Clonal variation in crown structure, absorbed photosynthetically active radiation and growth of loblolly pine and slash pine. Tree Physiol 27:421-430

Fabbrini F, Gaudet M, Bastien C, Zaina G, Harfouche A, Beritognolo I, Marron N, Morgante M, Scarascia-Mugnozza G, Sabatti M (2012) Phenotypic plasticity, QTL mapping and genomic characterization of bud set in black poplar. BMC Plant Biol 12:47. doi:10.1186/1471-2229-12-47

Farquhar GD, Ehleringer JR, Hubick KT (1989) Carbon Isotope Discrimination and Photosynthesis. Annu Rev Plant Phys 40:503-537. doi:10.1146/annurev.arplant.40.1.503

Flanagan LB, Johnsen KH (1995) Genetic variation in carbon isotope discrimination and Its relationship to growth under field conditions in full-sib families of *Picea mariana*. Can J Forest Res 25:39-47

Flint J, Mackay TF (2009) Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res 19:723-733. doi:10.1101/gr.086660.108

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31:3784-3788

Global climate change impacts in the United States (2009). Karl TR, Melillo JM, Peterson TC, (eds) Cambridge University Press, Cambridge

Gonzalez-Martinez SC, Krutovsky KV, Neale DB (2006) Forest-tree population genomics and adaptive evolution. New Phytol 170:227-238. doi:10.1111/j.1469-8137.2006.01686.x

González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in *Pinus taeda* L. I. wood property traits. Genetics 175:399-409. doi:10.1534/genetics.106.061127

Gugger PF, Fitz-Gibbon S, PellEgrini M, Sork VL (2016) Species-wide patterns of DNA methylation variation in *Quercus lobata* and their association with climate gradients. Mol Ecol 25:1665-1680. doi:10.1111/mec.13563

Harry DE, Temesgen B, Neale DB (1998) Codominant PCR-based markers for *Pinus taeda* developed from mapped cDNA clones. Theor Appl Genet 97:327-336. doi:DOI 10.1007/s001220050903

Havens K (1994) Clonal repeatability of in vitro pollen tube growth rates in *Oenothera organensis* (Onagraceae). Am J Bot 81:161-165. doi:10.2307/2445629

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33:54-78

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9:1322-1332. doi:10.1111/j.1755-0998.2009.02591.x

Hübner S, Smith HM, Hu W, Chan CK, Rihs H-P, Paschal BM, Raikhel NV, Jans DA (1999) Plant importin α binds nuclear localization sequences with high affinity and can mediate nuclear import independent of importin β. J Biol Chem 274:22610-22617

Huggett R, Wear DN, Li R, Coulston J, Liu S (2013) Forecasts of forest conditions. In: Wear DN, Greis JG (eds) The Southern Forest Futures Project: technical report. Gen. Tech.Rep. SRS-GTR-178. USDA-Forest Service, Southern Research Station, Asheville, NC, p 542

IPCC (2014a) Climate Change 2014: Sunthesis Report. Contribution of Working Groups I, II AND III to the Fifth Assessment Report of the intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland

IPCC (2014b) Climate Change 2014: Synthesis Report. Contribution of Working Groups I,II and III to the Fifth Assessment Report of the Intergovermental Panel on Climate Change[Core Writing Team,R.K. Pachauri and L.A. Meyer (eds.)]. IPCC,Geneva, Switzerland, Isik F (2014) Genomic selection in forest tree breeding: the concept and an outlook to the future. New Forest 45:379-401. doi:10.1007/s11056-014-9422-z

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9:166-177. doi:10.1093/bfgp/elq001

Jokela EJ, Martin TA, Vogel JG (2010) Twenty-five years of intensive forest management with southern pines: important lessons learned. J Forest 108:338-347

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723. doi:10.1534/genetics.107.080101

Kawakatsu T, Huang SS, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, Chen H, Dubin M, Lee CR, Wang C, Bemm F, Becker C, O'Neil R, O'Malley RC, Quarless DX, Genomes C, Schork NJ, Weigel D, Nordborg M, Ecker JR (2016) Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell 166:492-505. doi:10.1016/j.cell.2016.06.044

Kayihan GC, Huber DA, Morse AM, White TL, Davis JM (2005) Genetic dissection of fusiform rust and pitch canker disease traits in loblolly pine. Theor Appl Genet 110:948-958. doi:10.1007/s00122-004-1915-2

Keller I, Bensasson D, Nichols RA (2007) Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet 3:e22

Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, Hartigan J, Yandell M, Langley CH, Korf I, Neale DB (2010) The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11:420. doi:10.1186/1471-2164-11-420

Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas fir. Genetics 171:2029-2041. doi:10.1534/genetics.105.044420

Lambeth C, Hubert D (1997) Inheritance of branching and crown traits and their relationship to growth rate in loblolly pine. Paper presented at the 24th biennial southern forest tree improvement conference, Orlando, FL, June 9-12

Ledig FT, Guries RP, Bonefeld BA (1983) The relation of growth to heterozygosity in pitch pine. Evolution 37:1227-1238. doi:10.2307/2408843

Leister RT, Katagiri F (2000) A resistance gene product of the nucleotide binding site--leucine rich repeats class can form a complex with bacterial avirulence proteins in vivo. Plant J 22:345-354

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760. doi:10.1093/bioinformatics/btp324

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079. doi:10.1093/bioinformatics/btp352

Li MC, Zhu JJ, Zhang M (2013) Foliar carbon isotope discrimination and related traits along light gradients in two different functional-type tree species. Eur J Forest Res 132:815-824. doi:10.1007/s10342-013-0723-0

Lu M, Krutovsky KV, Nelson CD, Koralewski TE, Byram TD, Loopstra CA (2016) Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). BMC Genomics. doi:10.1186/s12864-016-3081-8

Maggs CA, Castilho R, Foltz D, Henzler C, Jolly MT, Kelly J, Olsen J, Perez KE, Stam W, Vainola R, Viard F, Wares J (2008) Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. Ecology 89:S108-122

Manna S (2015) An overview of pentatricopeptide repeat proteins and their applications. Biochimie 113:93-99. doi:10.1016/j.biochi.2015.04.004

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747-753. doi:10.1038/nature08494

McAdam SA, Sussmilch FC, Brodribb TJ (2016) Stomatal responses to vapour pressure deficit are regulated by high speed gene expression in angiosperms. Plant Cell Environ 39:485-491. doi:10.1111/pce.12633

McKeand SE, Jokela EJ, Huber DA, Byram TD, Allen HL, Li BL, Mullin TJ (2006) Performance of improved genotypes of loblolly pine across different soils, climates, and silvicultural inputs. Forest Ecol Manag 227:178-184. doi:10.1016/j.foreco.2006.02.016

Mclean RA, Sanders WL, Stroup WW (1991) A unified approach to mixed linear-models. Am Stat 45:54-64. doi:10.2307/2685241

Millar CI, Stephenson NL, Stephens SL (2007) Climate change and forests of the future: managing in the face of uncertainty. Ecol Appl 17:2145-2151

Namroud MC, Guillet-Claude C, Mackay J, Isabel N, Bousquet J (2010) Molecular evolution of regulatory genes in spruces from different species and continents: heterogeneous patterns of linkage disequilibrium and selection but correlated recent demographic changes. J Mol Evol 70:371-386. doi:10.1007/s00239-010-9335-1

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325-330. doi:10.1016/j.tplants.2004.05.006

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martinez-Garcia PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu LS, Gilbert D, Marcais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, deJong PJ, Yorke JA, Salzberg SL, Langley CH (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15:R59. doi:10.1186/gb-2014-15-3-r59

Neves L, Davis J, Barbazuk B, Kirst M Targeted sequencing in the loblolly pine (*Pinus taeda*) megagenome by exome capture. In: BMC Proceedings, 2011. vol Suppl 7. BioMed Central Ltd, p O48

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013a) A high-density gene map of loblolly pine (*Pinus taeda* L.) based on exome sequence capture genotyping. G3 (Bethesda):g3. 113.008714

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013b) Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J 75:146-156. doi:10.1111/tpj.12193

Noormets A, Gavazzi MJ, Mcnulty SG, Domec JC, Sun G, King JS, Chen JQ (2010) Response of carbon fluxes to drought in a coastal plain loblolly pine forest. Global Change Biol 16:272-287. doi:10.1111/j.1365-2486.2009.01928.x

Orians GH, Solbrig OT (1977) A cost-income model of leaves and roots with special reference to arid and semiarid areas. Am Nat 111:677-690

Palle SR, Seeve CM, Eckert AJ, Wegrzyn JL, Neale DB, Loopstra CA (2013) Association of loblolly pine xylem development gene expression with single-nucleotide polymorphisms. Tree Physiol 33:763-774. doi:10.1093/treephys/tpt054

Pasqualato S, Renault L, Cherfils J (2002) Arf, Arl, Arp and Sar proteins: a family of GTP‐binding proteins with a structural device for ‘front-back’ communication. EMBO Rep 3:1035-1041

Pessino M, Chabot ET, Giordano R, DeWalt RE (2014) Refugia and postglacial expansion of Acroneuria frisoni Stark & Brown (Plecoptera: Perlidae) in North America. Freshwater Science 33:232-249. doi:10.1086/675306

Plomion C, Bartholome J, Lesur I, Boury C, Rodriguez-Quilon I, Lagraulet H, Ehrenmann F, Bouffier L, Gion JM, Grivet D, de Miguel M, de Maria N, Cervera MT, Bagnoli F, Isik F, Vendramin GG, Gonzalez-Martinez SC (2016) High-density SNP assay development for genetic analysis in maritime pine (*Pinus pinaster*). Mol Ecol Resour 16:574-587. doi:10.1111/1755-0998.12464

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome-Us 5:92-102. doi:10.3835/plantgenome2012.05.0005

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Puig S (2014) Function and regulation of the plant COPT family of high-affinity copper transport proteins. Advances in Botany 2014, Article ID 476917, 9 pages. doi:10.1155/2014/476917

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575. doi:10.1086/519795

Quesada T, Gopal V, Cumbie WP, Eckert AJ, Wegrzyn JL, Neale DB, Goldfarb B, Huber DA, Casella G, Davis JM (2010) Association mapping of quantitative disease resistance in a natural population of loblolly pine (*Pinus taeda* L.). Genetics 186:677-686. doi:10.1534/genetics.110.117549

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842. doi:10.1093/bioinformatics/btq033

R Core Team (2015) R: A language and environment for statistical computing. Vienna, Austria; 2014. http://wwwR-projectorg

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573-589. doi:10.1534/genetics.114.164350

Remington DL, Whetten RW, Liu BH, O'Malley DM (1999) Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*. Theor Appl Genet 98:1279-1292. doi:DOI 10.1007/s001220051194

Renouard S, Tribalatc M-A, Lamblin F, Mongelard G, Fliniaux O, Corbin C, Marosevic D, Pilard S, Demailly H, Gutierrez L (2014) RNAi-mediated pinoresinol lariciresinol reductase gene silencing in flax (*Linum usitatissimum* L.) seed coat: Consequences on lignans and neolignans accumulation. J Plant Physiol 171:1372-1377. doi:10.1016/j.jplph.2014.06.005

Resende Jr M, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapalia D, Resende MD, Kirst M (2012) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments (vol 193, 617, 2012). New Phytol 193:617-624

Resende MF, Jr., Munoz P, Resende MD, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190:1503-1510. doi:10.1534/genetics.111.137026

Rodríguez-Quilón I, Santos-del-Blanco L, Grivet D, Jaramillo-Correa JP, Majada J, Vendramin GG, Alía R, González-Martínez SC (2015) Local effects drive heterozygosity–fitness correlations in an outcrossing long-lived tree. Proc Biol Sci 282:20152230. doi:10.1098/rspb.2015.2230

Ruiz-Lopez MJ, Ganan N, Godoy JA, Del Olmo A, Garde J, Espeso G, Vargas A, Martinez F, Roldan ER, Gomendio M (2012) Heterozygosity-fitness correlations and inbreeding depression in two critically endangered mammals. Conserv Biol 26:1121-1129. doi:10.1111/j.1523-1739.2012.01916.x

Schmidtling R (2001) Southern pine seed sources. In: Gen. Tech. Rep. SRS-44. U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC, p 25

Seeve CM (2010) Gene expression and association analyses of stress responses in loblolly pine (*Pinus taeda* L.). Dissertation, Texas A&M University

Shmulsky R, Jones PD (2011) Forest products and wood science. John Wiley & Sons, Hoboken

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J (2011) Fast, scalable generation of high‐quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. doi:10.1038/msb.2011.75

Slatkin M (2008) Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477-485. doi:10.1038/nrg2361

Smith WB, Miles PD, Perry CH, Pugh SA (2009) Forest resources of the United States, 2007. In: Gen. Tech. Rep. WO-78. U.S. Department of Agriculture, Forest Service, Wahsington Office, Washington, DC, p 336

Sun X, Sun C, Li Z, Hu Q, Han L, Luo H (2016) AsHSP17, a creeping bentgrass small heat shock protein modulates plant photosynthesis and ABA‐dependent and independent signalling to attenuate plant response to abiotic stress. Plant Cell Environ. doi:10.1111/pce.12683

Suren H, Hodgins KA, Yeaman S, Nurkowski KA, Smets P, Rieseberg LH, Aitken SN, Holliday JA (2016) Exome capture from the spruce and pine giga-genomes. Mol Ecol Resour. doi:10.1111/1755-0998.12570

Suzuki S, Umezawa T (2007) Biosynthesis of lignans and norlignans. J Wood Sci 53:273-284. doi:10.1007/s10086-007-0892-x

Tanenbaum ME, Macurek L, van der Vaart B, Galli M, Akhmanova A, Medema RH (2011) A complex of Kif18b and MCAK promotes microtubule depolymerization and is negatively regulated by Aurora kinases. Curr Biol 21:1356-1365. doi:10.1016/j.cub.2011.07.017

Thavamanikumar S, Southerton SG, Bossinger G, Thumma BR (2013) Dissection of complex traits in forest trees - opportunities for marker-assisted selection. Tree Genet Genomes 9:627-639. doi:10.1007/s11295-013-0594-z

Turner DP, Koerper GJ, Harmon ME, Lee JJ (1995) A Carbon budget for forests of the conterminous United States. Ecol Appl 5:421-436. doi:Doi 10.2307/1942033

Wegrzyn JL, Liechty JD, Stevens KA, Wu LS, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martinez-Garcia PJ, Holt C, Yandell M, Zimin AV, Yorke JA, Crepeau MW, Puiu D, Salzberg SL, Dejong PJ, Mockaitis K, Main D, Langley CH, Neale DB (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics 196:891-909. doi:10.1534/genetics.113.159996

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370. doi:Doi 10.2307/2408641

Wells OO, Switzer GL, Schmidtling RC (1991) Geographic variation in Mississippi loblolly pine and sweetgum. Silvae Genet 40:105-119

White TL, Adams WT, Neale DB (2007) Forest genetics. Cabi, Wallingford. doi:10.1079/9781845932855.0000

Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer Science & Business Media, http://had.co.nz/ggplot2/book

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859-1875. doi:10.1093/bioinformatics/bti310

Yakovlev I, Fossdal CG, Skroppa T, Olsen JE, Jahren AH, Johnsen O (2012) An adaptive epigenetic memory in conifers with important implications for seed production. Seed Sci Res 22:63-76. doi:10.1017/S0960258511000535

Yamaguchi-Shinozaki K, Shinozaki K (1993) The plant hormone abscisic acid mediates the drought-induced expression but not the seed-specific expression of rd22, a gene responsive to dehydration stress in *Arabidopsis thaliana*. Mol Gen Genet 238:17-25

Yang JT, Laymon RA, Goldstein LS (1989) A three-domain structure of kinesin heavy chain revealed by DNA sequence and microtubule binding analyses. Cell 56:879-889

Zgurski JM, Sharma R, Bolokoski DA, Schultz EA (2005) Asymmetric auxin response precedes asymmetric growth and differentiation of *asymmetric leaf1* and *asymmetric leaf2* Arabidopsis leaves. Plant Cell 17:77-91. doi:10.1105/tpc.104.026898

Zhang J, Singh A, Mueller DS, Singh AK (2015) Genome ‐ wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean. Plant J 84:1124-1136. doi:10.1111/tpj.13069

Zhang S (2015) Alaska’s on fire and It may make climate change even worse. Wired. http://www.wired.com/2015/07/alaskas-fire-may-make-climate-change-even-worse/. Accessed July 8, 2015

Zhang X, Gou M, Liu CJ (2013) Arabidopsis Kelch repeat F-box proteins regulate phenylpropanoid biosynthesis via controlling the turnover of phenylalanine ammonia-lyase. Plant Cell 25:4994-5010. doi:10.1105/tpc.113.119644

Zheng J, Chen F, Wang Z, Cao H, Li X, Deng X, Soppe WJ, Li Y, Liu Y (2012) A novel role for histone methyltransferase KYP/SUVH4 in the control of Arabidopsis primary seed dormancy. New Phytol 193:605-616. doi:10.1111/j.1469-8137.2011.03969.x

Zhu BL, Coleman GD (2001) The poplar bark storage protein gene (Bspa) promoter is responsive to photoperiod and nitrogen in transgenic poplar and active in floral tissues, immature seeds and germinating seeds of transgenic tobacco. Plant Mol Biol 46:383-394. doi:10.1023/A:1010600504740

Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marcais G, Puiu D, Roberts M, Wegrzyn JL, de Jong PJ, Neale DB, Salzberg SL, Yorke JA, Langley CH (2014) Sequencing and assembly of the 22-gb loblolly pine genome. Genetics 196:875-890. doi:10.1534/genetics.113.159715