Detecting the genetic basis of local adaptation in loblolly pine (*Pinus taeda* L.) using whole exome-wide genotyping and an integrative landscape genomics analysis approach

Mengmeng Lu[1, 2] | Carol A. Loopstra[1, 2] | Konstantin V. Krutovsky[1, 2, 3, 4, 5]

[1]Department of Ecosystem Science and Management, Texas A&M University, College Station, TX, USA

[2]Molecular and Environmental Plant Sciences Program, Texas A&M University, College Station, TX, USA

[3]Department of Forest Genetics and Forest Tree Breeding, Georg-August-University of Göttingen, Büsgenweg 2, 37077 Göttingen, Germany

[4]Laboratory of Population Genetics, N. I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina Str. 3, 119991 Moscow, Russia

[5]Laboratory of Forest Genomics, Genome Research and Education Center, Institute of Fundamental Biology and Biotechnology, Siberian Federal University, Akademgorodok 50a/2, 660036 Krasnoyarsk, Russia

Correspondence

Konstantin V. Krutovsky, Department of Forest Genetics and Forest Tree Breeding, Georg-August-University of Göttingen, Göttingen, Germany
Email: konstantin.krutovsky@forst.uni-goettingen.de

Present address: Mengmeng Lu, Department of Biological Sciences, University of Calgary, Calgary, AB, Canada

Abstract

In the Southern United States, the widely distributed loblolly pine contributes greatly to lumber and pulp production, as well as providing many important ecosystem services. Climate change may affect the productivity and range of loblolly pine. Nevertheless, we have insufficient knowledge of the adaptive potential and the genetics underlying the adaptability of loblolly pine. To address this, we tested the association of 2.8 million whole exome-based single nucleotide polymorphisms (SNPs) with climate and geographic variables, including temperature, precipitation, latitude, longitude and elevation data. Using an integrative landscape genomics approach by combining multiple environmental association and outlier detection analyses, we identified 611 SNPs associated with 56 climate and geographic variables. Longitude, maximum temperature of the warm months and monthly precipitation associated with most SNPs, indicating their importance and complexity in shaping the genetic variation in loblolly pine. Functions of candidate genes related to terpenoid synthesis, pathogen defense, transcription factors and abiotic stress response. We provided evidence that environment-associated SNPs also composed the genetic structure of adaptive phenotypic traits including height, diameter, metabolite levels and expression of genes. Our study promotes understanding of the genetic basis of local adaptation in loblolly pine, and provides promising tools for selecting genotypes adapted to local environments in a changing climate.

KEYWORDS

climate change, environmental association, loblolly pine, adaptability, outlier detection, SNP

1 | INTRODUCTION

Loblolly pine comprises 80% of the planted forestland and over one half of the standing volume in the Southern U.S. (Wear, Huggett, Li, Perryman, & Liu, 2013). The natural habitat of loblolly pine ranges from East Texas to central Florida and north to Southern New Jersey, demonstrating adaptability to various types of soil and growing conditions. Successful forest plantations rely on the selection of appropriate seed sources. The seed transfer guidelines for southern pines emphasize three key points: 1) low temperature to the north and low rainfall to the west limit the distribution of southern pines; 2) the annual average minimum temperature is the most important climate variable related to growth and survival; 3) for loblolly pine, seeds from east of the Mississippi River should not be used in the west because of the higher danger of losses due to droughts (Schmidtling, 2003).

As the climate changes, traditional seed selection guidelines may need to be adjusted to select for robust genotypes adapted to a changing climate scenario. An altered temperature and precipitation pattern threatens forests with droughts, fires and pathogen outbreaks, eventually leading to damage to the quality and yield of wood produced (Allen et al., 2010). Landscape genomics methods have been applied to explore the genetic basis of local adaptation in loblolly pine. The main objectives of these studies were to identify the environmental factors that have shaped the adaptive genetic variation and the gene variants that drive local adaptation (Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015; Sork et al., 2013). Eckert et al. (2010a) found five loci correlated with aridity and identified 24 loci as

F_{ST} outliers in loblolly pine. Eckert et al. (2010b) also found several well-supported loblolly pine SNPs associated with principal components corresponding to geography, temperature, growing degree-days, precipitation and aridity. Chhatre, Byram, Neale, Wegrzyn and Krutovsky (2013) detected SNPs as candidates for diversifying and balancing selection in natural and breeding loblolly pine populations in East Texas. Despite of the application of multiple methods, the size and complexity of conifer genomes limit the progress to further dissect the genetic basis of local adaptation.

In the current study, we aimed to discover more loci and genes with signatures of natural selection and incorporated phenotypic data into environmental adaption analyses to improve insight. We have discovered 2.8 million SNPs using whole exome sequencing from a clonally propagated association mapping loblolly pine population (Lu et al., 2016; Lu et al., 2017; Lu, Seeve, Loopstra, & Krutovsky, 2018). This population represented diverged ecophysiological regions across 12 states in the Southern U.S., extending from Texas to Virginia. Loblolly pine populations have shown adaptation to environment based on the geographic distributions of traits. For example, loblolly pines from west of the Mississippi River are slower growing, but more resistant to fusiform rust, drought and crowding than trees from east of the Mississippi River (Schmidtling, 1988; Schmidtling & Froelich, 1993; Wells, 1985). We examined associations of 2.8 million whole exome-based SNPs with climate and geographic variables in 328 loblolly pine trees using a landscape genomics approach integrating multiple analysis methods. We detected SNPs associated with both adaptive phenotypic traits and climate/geographic variables, and identified candidate genes that contribute to local adaptation in loblolly pine. The results can help determine how selection affects the genetic architecture of adaptive traits. The identified loci and genes can contribute to rapid selection of genotypes with adaptive potential to climate change.

2 | MATERIALS AND METHODS

2.1 | Genotypic data

The loblolly pine population used in this study and the process of obtaining genotyping data were previously described in Lu et al. (2017). Briefly, we analyzed 328 trees with a clearly known origin. They were divided into 3 regions as described by Schmidtling (2001): 1) 304 trees representing the eastern region, including states east of the Mississippi River; 2) 13 trees representing the western region, including the states of Arkansas and Louisiana; 3) 11 trees representing the far west region, including the states of Texas and Oklahoma.

2.2 | Climate and geographic data

Climate and geographic data for each tree in the population were the same as in Eckert et al. (2010a). The data were originally gathered from the WORLDCLIM 2.5-min geographical information system (GIS) layer using Diva-GIS v.5.4 (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005). The dataset contained a total of 58 variables, including latitude, longitude, elevation, average minimum and maximum temperature for each month, average precipitation for each month, and 19 bioclimatic variables. The bioclimatic variables are summary statistics of precipitation and temperature. For example, BIO1 represents annual mean temperature, and BIO12 represents annual precipitation. Details of these 19 bioclimatic variables are presented in Table S1. The JMP Pro 12 statistical software (SAS Institute, Cary, NC) was used to display the variation of climate variables across the counties. A principle component analysis (PCA) of these variables was carried out using the *prcomp* function in R (R_Core_Team, 2017). The PCA was visualized by the R package *ggbiplot* (https://github.com/vqv/ggbiplot/tree/experimental).

2.3 | Environmental associations and outlier analyses

Multiple approaches were employed to discover the loci associated with climate and geographic variables. The process is schematically summarized in Figure 1. Specifically, we studied association between 2.8 million SNPs and climate/geographic variables using TASSEL 5.0 (Bradbury et al., 2007). The procedure was the same as previously described in Lu et al. (2017). In addition, two outlier detection methods were employed to detect loci under selection and potentially involved in local adaption. One method is the spatial ancestry analysis (SPA), which identifies SNPs with significant gradients in allele frequency (Yang, Novembre, Eskin, & Halperin, 2012). The geographical location (longitude and latitude) information for each tree was supplied as the “--location-input”. SNPs with SPA scores above the 99.9% percentile were considered as outliers. Another outlier detection method was implemented by the OutFLANK software (Whitlock & Lotterhos, 2015). It infers the F_{ST} distribution for a large set of loci and identifies the loci that may contribute to a significant local differentiation and potential adaptation. A *Q*-value of 0.05 was applied to detect outliers. Following the program recommendation, 1,323,910 SNPs with a minor allele frequency (MAF) >= 0.05 were used for the SPA and OutFLANK analyses.

We used multivariate analysis to identify the significance of climate in structuring genetic diversity among the outlier SNPs. The multivariate relationships were examined using the redundancy analysis (RDA) implemented in the R package *vegan* (Oksanen et al., 2017; R_Core_Team, 2017). We estimated the proportion of SNP variation explained by only climate variables using a partial redundancy analysis (pRDA), in which the effects of climate variables were conditioned on the effects of geography. Statistical significance of the pRDA estimates was assessed using a permutation-based analysis of variance (ANOVA).

Association of the outlier loci with climate and geographic variables was analyzed using the Samβada software (Stucki et al., 2017). This software is based on the logistic regression model and assesses whether the allelic variation correlates with specific environmental variables. Spatial association due to population structure is accounted for by measuring indices of spatial autocorrelation. In this study, the parameters for Samβada analysis were set up as: spatial autocorrelation was measured along longitude and latitude using spherical coordinate and 20 nearest neighbors; both global and local autocorrelation of loci were included, and the significance was assessed with 1,000 permutations. The detection of selection signatures was based on univariate models and the threshold for screening significant models was set to 1%.

We searched for SNPs associated with both adaptive phenotypic traits and climate/geographic variables to better understand how selection pressures shape the genetic structures underlying local adaptation. Using the same SNP set and population, we previously found SNP associations with such adaptive phenotypic traits as specific leaf area, branch angle, height, diameter, crown width, carbon isotope discrimination, and nitrogen content (Lu et al., 2017). We also found SNP associations with metabolite levels and expression of wood development- and stress resistance-related genes (Lu et al., 2018). In this study, we focused on SNPs that have associations with both climate/geographic variables and adaptive phenotypic traits. The JMP Pro 12 statistical software (SAS Institute, Cary, NC) was employed to display the variation of climate/geographic variables, genotypes, and phenotypic traits.

The annotation for genes that contain identified SNPs was obtained from loblolly pine gene annotation files available on https://treegenesdb.org/FTP/Genomes/Pita/v1.01/annotation/ (Wegrzyn et al., 2014). The regulatory sequences including promoters, enhancers and silencers have not yet been identified. 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 is 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 VCFtools software (Danecek et al., 2011) was used to calculate the MAF.

3 | RESULTS

3.1 | Climate variation in the loblolly pine natural range

Among the counties of origin for the studied trees, the annual mean temperature (BIO1) demonstrated a decreasing trend from South to North (Figure 2a). The annual precipitation (BIO12) was higher in Louisiana, Mississippi and Alabama than in other regions (Figure 2b). Maximum temperature of the warmest month (BIO5) and mean temperature of the driest quarter (BIO9) were higher in the western and far west regions (Figure S1). Mean temperature of the wettest quarter (BIO8), precipitation seasonality (BIO15), and precipitation of wettest and warmest quarter (BIO16 & BIO18) were higher in the eastern region. Precipitation of the coldest quarter (BIO19), driest month (BIO14), and driest quarter (BIO17) were higher in Louisiana, Mississippi and Alabama compared with other states. Along South to North, minimum temperature of the coldest month (BIO6) and mean temperatures of the warmest and coldest quarters (BIO10 & BIO11) decreased, while temperature seasonality (BIO4) and annual temperature range (BIO7) increased. The PCA of the climate variables showed different climate conditions among the counties of origin for the studied trees (Figure 3). The first PC was mainly correlated with temperature variables, explaining 62.6% of the variation of the climate variables. The second PC was mainly correlated with precipitation variables, explaining 21.4% of the variation of the climate variables.

3.2 | SNPs associated with climate and geographic variables

We identified 503 associations, including 49 climate/geographic variables and 293 SNPs (Table S2). Among them, 297 associations involved temperature variables, 174 - precipitation variables, 21 - elevation, and 11 - latitude. The MAF of the identified SNPs were between 0.01 and 0.5 with a median of 0.02. Among the 293 SNPs, 199 were in 195 annotated genes. Specifically, 3 SNPs (2%) were in 3’ regulatory sequences (3’ RS), 9 (4%) in 5’ RS, 118 (59%) in coding sequences (CDS), 59 (29%) in introns, 5 (3%) in 5’ untranslated regions (5’ UTR), and 5 (3%) in 3’ UTR. The remaining SNPs were in unclassified or intergenic regions. Most identified SNPs were associated with multiple variables. For example, the SNP tscaffold3881_229913 was associated with latitude, 3 precipitation variables, and 25 temperature variables. This SNP resides in the CDS of a gene encoding EARLY FLOWERING 3-like protein, which is a circadian clock protein playing key roles in adaptation of plants to diurnal environmental conditions.

3.3 | Outlier SNPs

We found that 1,324 SNPs showed large gradients in allele frequency based on the SPA analysis (Table S3). Among them, 1,099 SNPs resided in 381 annotated genes. Specifically, 43 SNPs (4%) resided in 3’ RS, 68 (6%) in 5’ RS, 548 (50%) in CDS, 380 (35%) in introns, 14 (1%) in 5’ UTR, and 46 (4%) in 3’UTR. The other SNPs resided in unclassified or intergenic regions. The annotated genes PITA_000021128 and PITA_000021125 contained the most outlier SNPs, 38 and 27, respectively. These two genes encode the ent-copalyl diphosphate synthase, and the abietadienol/abietadienal oxidase-like protein, respectively. Both genes participate in terpenoid synthesis and contribute to conifer defense against herbivores and pathogens.

We also identified 242 SNP outliers using the OutFLANK software (Table S4). Among them, 189 SNPs resided in 128 annotated genes. Specifically, 8 SNPs (4%) resided in 3’ RS, 11 (6%) in 5’ RS, 120 (64%) in CDS, 44 (23%) in introns, 2 (1%) in 5’ UTR, and 4 (2%) in 3’UTR. The remaining SNPs resided in unclassified or intergenic regions. The annotated genes PITA_000091177, PITA_000064023, and PITA_000040532 contained the most outlier SNPs. These three genes encode a LRR receptor-like serine/threonine-protein kinase, a bHLH transcription factor, and a protein of unknown function.

We found 33 loci identified by both SPA and OutFLANK software (Table S5). The MAFs of these 33 loci ranged between 0.06 and 0.47 with a median of 0.21. These 33 loci resided in 12 annotated genes encoding proteins that include the leucine-rich repeat receptor-like serine/threonine-protein kinase, the bHLH transcription factor, oxidoreductase, and an EARLY FLOWERING 3-like protein.

3.4 | Multivariate analyses of the identified SNP outliers

The pRDA model confirmed that the outlier SNPs are significantly correlated ($P < 0.001$) with climate and geography. Climate and geography alone explained 50% and 1% of the SNP outliers' variance, respectively. However, the remaining proportion of variance was rather large due to the joint effect of climate and geography demonstrating their interactive influence on the SNP variation. We plotted a pRDA biplot graph to visualize important climate and geographic variables shaping the genetic variation (Figure S2). In general, precipitation variables dominated the pRDA axis 1. The most important variables in explaining variation of SNP outliers along the pRDA axis 1 were average precipitation in January, February, March, April and December, precipitation of the driest quarter (BIO17), mean temperature of the wettest quarter (BIO8), mean diurnal range (BIO2), and precipitation of the driest month (BIO14).

3.5 | Outlier SNPs associated with climate and geographic variables

We identified 1,790 associations between 323 SNP outliers and 47 climate/geographic variables using the Samβada software (Table S6). Among them, 963 associations were related to temperature, 476 to precipitation, 41 to latitude and 310 to longitude. The outlier SNPs associated with environment had MAFs between 0.05 and 0.49 with a median of 0.21, residing in 250 annotated genes.

Taken together, we identified 611 unique SNPs associated with 56 climate and geographic variables ("environmental SNPs" - envSNPs) using either the TASSEL or Samβada software.

Only two variables, precipitation seasonality (BIO15) and precipitation of the driest quarter (BIO17) were not found to be associated with any SNP. Of the other variables, longitude was associated with the most SNPs (310), followed by maximum temperature of August (206), precipitation of May (168), maximum temperature of July (159), maximum temperature of the warmest month (BIO5) (155), precipitation of November (107), maximum temperature of September (76), mean temperature of the driest quarter (BIO9) (76), precipitation of December (67), maximum temperature of June (59), and mean temperature of the warmest quarter (BIO10) (59) (Figure 4).

We categorized genes containing the 611 envSNPs into four main functional groups: 1) terpenoid synthesis, 2) pathogen and disease defense, 3) transcription factors, and 4) abiotic stress response (Tables 1 and S7). Among the 611 envSNPs, five SNPs (scaffold10517.2_56785, scaffold674735_1427, scaffold721455_39357, tscaffold3881_229913, tscaffold551_336950) were detected by both software. They resided in the following four annotated genes: PITA_000048497, PITA_000060878, PITA_000004436, and PITAhm_001489, which encode an abietadienol/abietadienal oxidase-like protein, a myrcene synthase or terpene synthase metal-binding domain protein, an EARLY FLOWERING 3-like protein, and a DEAD/DEAH box helicase domain protein.

3.6 | SNPs associated with both climate/geographic variables and adaptive phenotypic traits

We identified five envSNPs associated with both height and diameter, 10 with height only, 114 with 27 metabolite levels, and 242 with expression levels of 47 genes (Tables 2, S8 and S9). For example, 54 envSNPs associated with arachidic acid levels, and more than 60 envSNPs associated with the expression levels of *ANR* and *NCED* genes.

We combined genomic, phenotypic and climate/geographic data to analyze adaptive genetic variation. For example, we found the envSNP scaffold10517.2_56785 (identified by both association and outlier detection methods) correlated with expression levels of the *ANR* and *NCED* genes. The expression levels of these two genes also correlated with precipitation of May (Figure 5a). The *ANR* gene encodes an anthocyanidin reductase, which is important for the biosynthesis of condensed tannins (Xie, Sharma, Paiva, Ferreira, & Dixon, 2003). The *NCED* gene encodes a 9-*cis* epoxycarotenoid dioxygenase, which prepares precursors for synthesis of abscisic acid (ABA) (Tan et al., 2003). ABA is a key regulator of seed development, root growth, stomatal aperture and plant responses to water stress. The envSNP scaffold10517.2_56785 resided in a gene encoding an abietadienol/abietadienal oxidase-like protein, which is a multifunctional and multisubstrate cytochrome P450 monooxygenase that contributes to conifer defense by generating an enormous structural diversity of plant terpenoid secondary metabolites (Ro, Arimura, Lau, Piers, & Bohlmann, 2005). Individuals with the AA genotype tended to have low expression of the *ANR* gene and high expression of the *NCED* gene (Figure 5b). They were common in counties with low precipitation in May. On the contrary, individuals with the GG genotype had high expression of the *ANR* gene, and low expression of the *NCED* gene. They were common in counties with high precipitation in May. Individuals with the AG genotype were common in counties with medium precipitation in May, and the expression of the *ANR* and *NCED* genes did not differ much from the individuals with the AA genotypes. Precipitation in May positively correlated with the *ANR* gene expression level (r = 0.4, P < 0.0001) and negatively correlated with the *NCED* gene expression level (r = -0.2, P=0.0005).

4 | DISCUSSION

We identified 611 envSNPs associated with 56 climate and geographic variables. Longitude, maximum temperature of the warm months and monthly precipitation associated with most envSNPs. The identified envSNPs resided in genes related to terpenoid synthesis, pathogen and disease defense, transcription factors and abiotic stress response. We also found that some envSNPs composed the genetic structure of adaptive phenotypic traits including height, diameter, metabolite levels and expression of genes.

4.1 | Comparison of multiple analysis methods

Combining environmental association analyses with outlier detection methods is a desirable way to reduce the rate of false positives and assess the relevance of findings in landscape genomic research (Le Corre & Kremer, 2012; Rellstab et al., 2015), but each method has its strengths and weaknesses. TASSEL exploits the genomic diversity at a very high resolution, hence it is sensitive for detecting associations even for SNPs with low MAFs. In this study, among the 293 envSNPs that demonstrated significant associations with climate and geographic variables detected by TASSEL, 72% had a MAF less than 0.05. Associations could be due to linkage disequilibrium with the functional loci and hence not directly involved in environmental adaptation. The SPA and OutFLANK software detect SNPs under strong selection. To apply these two methods, loci with low MAFs (< 0.05) were removed due to a probable high sampling variance, which may negatively affect the power of models. This is especially critical for OutFLANK, because the distribution of F_{ST} for loci with low MAFs is very different from that for loci with more equal allele frequencies (Whitlock & Lotterhos, 2015). The MAFs of SNPs detected by SPA ranged from 0.06 to 0.5 with a median of 0.36. The MAFs of SNPs detected by OutFLANK ranged from 0.05 to 0.47 with a median of 0.07. Since most adaptation related traits are polygenic with small allele frequency changes at many loci (Le Corre & Kremer, 2012; Mackay, Stone, & Ayroles, 2009), SPA and OutFLANK would miss those loci under weak selection. Additionally, SPA and OutFLANK cannot identify the specific factors that drive selection. To further determine the selective factors, the Samβada software was applied to associate climate and geographic variables with SNP outliers while taking into account spatial autocorrelation. The Bonferroni correction implemented in the current Samβada software may be overly-conservative and may result in overlooking potentially adaptive loci (Stucki et al., 2017). We applied the multivariate approach RDA to examine the relationship between climate/geographic variables and genetic variation of the outlier SNPs. We identified precipitation factors as the important drivers for local adaption. However, the joint effect of climate and geography due to collinearity comprises 49% of the SNP outlier variance. The strong pattern of collinearity could skew the results (Rellstab et al., 2015).

The overlap rate among the SNPs detected by different software was relatively low. Among the 1324 and 242 SNP outliers detected by SPA and OutFLANK, respectively, only 33 SNPs were the same. Among the 293 and 323 envSNPs identified by TASSEL and Samβada, respectively, only 5 envSNPs were the same. Different assumptions and models applied in different software cause the relatively low numbers of consensus envSNPs. The low consistency across different genome scan methods was also reported previously (de Villemereuil, Frichot, Bazin, François, & Gaggiotti, 2014). There is no single widely accepted statistical approach (Rellstab et al., 2015). Integrating multiple methods and compiling all possible results can provide more reliable information for downstream analyses. Follow-ups are needed to validate the detected adaptive loci and genes using independent populations, knockout mutants, common garden, and reciprocal transplant experiments (Rellstab et al., 2015).

4.2 | Evidence of selection by environment

The identified SNP-environment associations helped us recognize the climate and geography variables that have shaped the genetic variation. We found that longitude, maximum temperature of the warm months and monthly precipitation were variables associated with the most envSNPs (Figure 4). They acted as selective factors driving loblolly pine local adaptation. Although the seed transfer guidelines advised the yearly average minimum temperature as the most important climate variable for southern pines (Schmidtling, 2003), the current study highlights the importance and complexity of maximum temperature of the warm months and monthly precipitation in shaping the genetic variation underlying loblolly pine adaptability. A significant increase in the number of consecutive days exceeding 35°C (a metric used as a measure of heat waves) and a decline in the net water supply availability are expected over the next decades, particularly in the western part of the loblolly pine range (Kunkel et al., 2013; Sun et al., 2013). In a rapid climate change scenario, if adaptation of loblolly pine cannot match the increased heat and drought conditions, the productivity and thus the economic and ecological profits will be greatly damaged. Selecting and planting genotypes adapted to the changing climate may reduce losses in loblolly pine plantations.

The identified candidate genes directly or indirectly related to abiotic or biotic stress response, including four functional groups: 1) terpenoid synthesis, 2) pathogen and disease defense, 3) transcription factors, and 4) abiotic stress response (Tables 1 and S7). For example, genes encoding the myrcene synthase and cytochrome P450 are in the terpenoid biosynthesis pathway. Terpenes offer chemical defense against herbivores and pathogens in conifers. The gene encoding a LRR receptor-like serine/threonine-protein kinase is related to pathogen and disease resistance. The transcription factors bHLH and MADS-box regulate downstream defensive and developmental reactions. Other genes are related to responses to abiotic stresses, including stresses from UV, salt, drought, nitrogen, cold, heat, oxidation and wounding. These stress response genes contribute to the genetic structure of loblolly pine adaptability, conferring mitigation and adaptation potential in diverse environments. Five genes related to loblolly pine adaptability and detected in the current study were also reported earlier in Eckert et al. (2010a). These consistently detected genes encode the MATE efflux family protein, a methyltransferase, a translation initiation factor, an ubiquitin, and an auxin responsive protein. They are associated with multiple climate and geographic variables including longitude, monthly precipitation and average maximum monthly temperature. For example, the gene encoding the MATE efflux family protein was previously identified to correlate with aridity (Eckert et al., 2010a). In the current study, this gene was found to be associated with average maximum temperature in February and March, precipitation in January, February, April, June, November and December, mean temperature of the driest quarter (BIO9), annual precipitation (BIO12) and precipitation of the coldest quarter (BIO19). The MATE efflux family proteins play important roles in a wide range of biological processes, such as transporting secondary metabolites, regulating disease resistance and detoxifying toxic compounds (Liu, Li, Wang, Gai, & Li, 2016). These consistently detected genes are strong candidates underlying loblolly pine adaptability.

Combining environmental association analyses with dissection of phenotypic traits can greatly improve our understanding of the genetic basis of local adaptation. Talbot et al. (2017) reported that loci with local adaptation signatures in loblolly pine were also linked to gene expression traits for lignin development and whole-plant traits. In our study, more associations between loci with local adaption signatures and adaptive phenotypic traits were detected due to the application of 2.8 million SNPs. The loci with local adaption signatures correlated with height, diameter, metabolite levels, and expression of genes. These results indicate that genes underlying adaptive phenotypic traits are likely involved in adaptability to the environment. These candidate genes need to be further tested in validation populations located in different environments.

5 | CONCLUSION

We identified 611 SNPs associated with 56 climate and geographic variables using an integrative landscape genomics approach by combining association analyses with outlier detection analyses. Longitude, maximum temperature of the warm months and monthly precipitation associated with most SNPs, indicating their importance and complexity in shaping the genetic variation underlying loblolly pine adaptability. The identified SNPs resided in genes related to terpenoid synthesis, pathogen and disease defense, transcription factors and abiotic stress response. We provided evidence that environment-associated SNPs (envSNPs) also composed the genetic structure of adaptive phenotypic traits including height, diameter, metabolite levels and expression of genes. The climate trend in the loblolly pine range -- increasing heat and drought -- pose challenges for breeding loblolly pine adapted to the planting environment. Our study provides envSNPs and candidate genes to facilitate elucidation of the genetic architecture of environmental adaptation in loblolly pine. The knowledge can be applied in breeding loblolly pine trees adapted to the future local environment.

ACKNOWLEDGEMENTS

We thank the Allele Discovery of Economic Pine Traits 2 (ADEPT 2) project (National Science Foundation Grant DBI-0501763) for developing the population, the PineRefSeq Project (USDA National Institute of Food and Agriculture, Award #2011-67009-30030) for providing the draft loblolly pine reference sequences, exon annotation and bioinformatics assistance, and the Texas A&M Institute for Genome Sciences and Society (TIGSS) for providing computational resources and system administration support for the TIGSS HPC Cluster. This study was funded 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.

DATA ACCESSIBILITY

All the data generated during this study were attached in the supplementary document. The Illumina HiSeq short read sequences that were used to detect the SNPs are deposited in the Sequence Read Archive (SRA) (accession number SRP075363; https://www.ncbi.nlm.nih.gov/sra).

AUTHOR CONTRIBUTIONS

C.A.L and K.V.K. conceived idea, designed the study, obtained the funding, coordinated the laboratory and field work, and assisted with editing the manuscript. ML performed the sample collection, data generation and analyses, and wrote the draft manuscript. All authors read and approved the final manuscript.

DISCLOSURE DECLARATION

The authors declare no competing interest.

ORCID

Konstantin V. Krutovsky http://orcid.org/0000-0002-8819-7084

Mengmeng Lu https://orcid.org/0000-0001-5023-3759

REFERENCES

Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., ... Cobb, N. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest Ecology and Management, 259*, 660-684. doi: 10.1016/j.foreco.2009.09.001

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics, 23*, 2633-2635. doi: 10.1093/bioinformatics/btm308

Chhatre, V. E., Byram, T. D., Neale, D. B., Wegrzyn, J. L., & Krutovsky, K. V. (2013). Genetic structure and association mapping of adaptive and selective traits in the east Texas loblolly pine (*Pinus taeda* L.) breeding populations. *Tree Genetics & Genomes, 9*, 1161-1178. doi: 10.1007/s11295-013-0624-x

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... 1000_Genomes_Project_Analysis_Group (2011). The variant call format and

VCFtools. *Bioinformatics, 27*, 2156-2158. doi: 10.1093/bioinformatics/btr330

de Villemereuil, P., Frichot, É., Bazin, É., François, O., & Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology, 23*, 2006-2019. doi: 10.1111/mec.12705

Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., Gonzάlez - Martίnez, S. C., & Neale, D. B. (2010a). 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, A. J., Bower, A. D., Gonzάlez - Martίnez, S. C., Wegrzyn, J. L., Coop, G., & Neale, D. B. (2010b). Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology, 19*, 3789-3805. doi: 10.1111/j.1365-294X.2010.04698.x

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology, 25*, 1965-1978. doi: 10.1002/joc.1276

Kunkel, K. E., Stevens, L. E., Stevens, S. E., Sun, L., Janssen, E., Wuebbles, D., ... Dobson, J. G. (2013). *NOAA technical report NESDIS 142-2. Regional climate trends and scenarios for the U.S. national climate assessment. Part 2. Climate of the Southeast U.S.* Washington, DC: National Environmental Satellite, Data, and Information Service.

Le Corre, V., & Kremer, A. (2012). The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology, 21*, 1548-1566. doi: 10.1111/j.1365-294X.2012.05479.x

Liu, J., Li, Y., Wang, W., Gai, J., & Li, Y. (2016). Genome-wide analysis of MATE transporters and expression patterns of a subgroup of *MATE* genes in response to aluminum toxicity in soybean. *BMC Genomics, 17*, 223. doi: 10.1186/s12864-016-2559-8

Lu, M., Krutovsky, K. V., Nelson, C. D., Koralewski, T. E., Byram, T. D., & Loopstra, C. A. (2016). Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). *BMC Genomics, 17*, 730. doi: 10.1186/s12864-016-3081-8

Lu, M., Krutovsky, K. V., Nelson, C. D., West, J. B., Reilly, N. A., & Loopstra, C. A. (2017). Association genetics of growth and adaptive traits in loblolly pine (*Pinus taeda* L.) using whole-exome-discovered polymorphisms. *Tree Genetics & Genomes, 13*, 57. doi: 10.1007/s11295-017-1140-1

Lu, M., Seeve, C. M., Loopstra, C. A., & Krutovsky, K. V. (2018). Exploring the genetic basis of gene transcript abundance and metabolite levels in loblolly pine (*Pinus taeda* L.) using association mapping and network construction. *BMC genetics, 19*, 100. doi: 10.1186/s12863-018-0687-7

Mackay, T. F., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics, 10*, 565-577. doi: 10.1038/nrg2612

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2017). vegan: Community Ecology Package.R package version 2.4-2.

https://CRAN.R-project.org/package=vegan

R_Core_Team (2017). R: A language and environment for statistical computing. https://www.R-project.org/

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology, 24*, 4348-4370. doi: 10.1111/mec.13322

Ro, D. K., Arimura, G. I., Lau, S. Y., Piers, E., & Bohlmann, J. (2005). Loblolly pine abietadienol/abietadienal oxidase *PtAO* (CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 8060-8065. doi: 10.1073/pnas.0500825102

Schmidtling, R. C. (1988). Racial variation in self-thinning trajectories in loblolly pines. In *Proceedings of the IUFRO Conference,* pp. 611-618.

Schmidtling, R. C., & Froelich, R. C. (1993). Thirty-seven year performance of loblolly pine seed sources in eastern Maryland. *Forest Science*, 39, 706-721.

Schmidtling, R. C. (2001). *General technical report SRS-44*. *Southern pine seed sources.* Asheville, NC: USDA Forest Service, Southern Research Station.

Schmidtling, R. C. (2003). Determining seed transfer guidelines for southern pines. In L. E. Riley, R. K. Dumroese & T. D. Landis, technical coordinators, *National proceedings: Forest and conservation nursery associations-2002. Proceedings RMRS-P-28*. (pp. 8-11). Ogden, UT: USDA Forest Service, Rocky Mountain Research Station.

Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., & Neale, D. B. (2013). Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics & Genomes, 9*, 901-911. doi: 10.1007/s11295-013-0596-x

Stucki, S., Orozco - terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., ... Joost, S. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources, 17*, 1072-1089. doi: 10.1111/1755-0998.12629

Sun, G., Caldwell, P. V., McNulty, S. G., Georgakakos, A. P., Arumugam, S., Cruise, J., ... Marion, D. A. (2013). *NCA Southeast technical report*. *Impacts of climate change and variability on water resources in the Southeast USA.* Asheville, NC: USDA Forest Service, Southern Research Station.

Talbot, B., Chen, T.-W., Zimmerman, S., Joost, S., Eckert, A. J., Crow, T. M., ... Manel, S. (2017). Combining genotype, phenotype, and environment to infer potential candidate genes. *Journal of Heredity, 108*, 207-216. doi: 10.1093/jhered/esw077

Tan, B. C., Joseph, L. M., Deng, W. T., Liu, L., Li, Q. B., Cline, K., & McCarty, D. R. (2003). Molecular characterization of the *Arabidopsis* 9-*cis* epoxycarotenoid dioxygenase gene family. *The Plant Journal, 35*, 44-56. doi: 10.1046/j.1365-313X.2003.01786.x

Wear, D. N., Huggett, R., Li, R., Perryman, B., & Liu, S. (2013). *Gen.Tech.Rep.SRS-GTR-170*. *Forecasts of forest conditions in regions of the United States under future scenarios:*

a technical document supporting the Forest Service 2012 RPA Assessment. Asheville, NC: USDA Forest Service, Southern Research Station.

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., ... Neale, D. B. (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

Wells, O. O. (1985). Use of Livingston Parish, Louisiana loblolly pine by forest products industries in the Southeast. *Southern Journal of Applied Forestry, 9*, 180-185. doi: 10.1093/sjaf/9.3.180

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. *The American Naturalist, 186*, S24-S36. doi: 10.1086/682949

Xie, D.-Y., Sharma, S. B., Paiva, N. L., Ferreira, D., & Dixon, R. A. (2003). Role of anthocyanidin reductase, encoded by BANYULS in plant flavonoid biosynthesis. *Science, 299*, 396-399. doi: 10.1126/science.1078540

Yang, W.-Y., Novembre, J., Eskin, E., & Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics, 44*, 725-731. doi: 10.1038/ng.2285

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information section for this article.