

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

4 Mengmeng Lu^{1,2} | Carol A. Loopstra^{1,2} | Konstantin V. Krutovsky^{1, 2, 3, 4, 5}

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

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

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

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

13 ⁵Laboratory of Forest Genomics, Genome Research and Education Center, Institute of
14 Fundamental Biology and Biotechnology, Siberian Federal University, Akademgorodok
15 50a/2, 660036 Krasnoyarsk, Russia

16 **Correspondence**

17 Konstantin V. Krutovsky, Department of Forest Genetics and Forest Tree Breeding, Georg-
18 August-University of Göttingen, Göttingen, Germany

19 Email: konstantin.krutovsky@forst.uni-goettingen.de

20

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

Calgary, AB, Canada

21 **Abstract**

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

39 **KEYWORDS**

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

41

42 **1 | INTRODUCTION**

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

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

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

69 In the current study, we aimed to discover more loci and genes with signatures of natural
70 selection and incorporated phenotypic data into environmental adaption analyses to improve
71 insight. We have discovered 2.8 million SNPs using whole exome sequencing from a clonally
72 propagated association mapping loblolly pine population (Lu et al., 2016; Lu et al., 2017; Lu,
73 Seeve, Loopstra, & Krutovsky, 2018). This population represented diverged ecophysiological
74 regions across 12 states in the Southern U.S., extending from Texas to Virginia. Loblolly pine
75 populations have shown adaptation to environment based on the geographic distributions of
76 traits. For example, loblolly pines from west of the Mississippi River are slower growing, but
77 more resistant to fusiform rust, drought and crowding than trees from east of the Mississippi
78 River (Schmidtling, 1988; Schmidtling & Froelich, 1993; Wells, 1985). We examined
79 associations of 2.8 million whole exome-based SNPs with climate and geographic variables in
80 328 loblolly pine trees using a landscape genomics approach integrating multiple analysis
81 methods. We detected SNPs associated with both adaptive phenotypic traits and

82 climate/geographic variables, and identified candidate genes that contribute to local
83 adaptation in loblolly pine. The results can help determine how selection affects the genetic
84 architecture of adaptive traits. The identified loci and genes can contribute to rapid selection
85 of genotypes with adaptive potential to climate change.

86 **2 | MATERIALS AND METHODS**

87 **2.1 | Genotypic data**

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

94 **2.2 | Climate and geographic data**

95 Climate and geographic data for each tree in the population were the same as in Eckert et al.
96 (2010a). The data were originally gathered from the WORLDCLIM 2.5-min geographical
97 information system (GIS) layer using Diva-GIS v.5.4 (Hijmans, Cameron, Parra, Jones, &
98 Jarvis, 2005). The dataset contained a total of 58 variables, including latitude, longitude,
99 elevation, average minimum and maximum temperature for each month, average precipitation

100 for each month, and 19 bioclimatic variables. The bioclimatic variables are summary statistics
101 of precipitation and temperature. For example, BIO1 represents annual mean temperature, and
102 BIO12 represents annual precipitation. Details of these 19 bioclimatic variables are presented
103 in Table S1. The JMP Pro 12 statistical software (SAS Institute, Cary, NC) was used to
104 display the variation of climate variables across the counties. A principle component analysis
105 (PCA) of these variables was carried out using the *prcomp* function in R (R_Core_Team,
106 2017). The PCA was visualized by the R package *ggbiplot*
107 (<https://github.com/vqv/ggbiplot/tree/experimental>).

108 **2.3 | Environmental associations and outlier analyses**

109 Multiple approaches were employed to discover the loci associated with climate and
110 geographic variables. The process is schematically summarized in Figure 1. Specifically, we
111 studied association between 2.8 million SNPs and climate/geographic variables using
112 TASSEL 5.0 (Bradbury et al., 2007). The procedure was the same as previously described in
113 Lu et al. (2017). In addition, two outlier detection methods were employed to detect loci
114 under selection and potentially involved in local adaptation. One method is the spatial ancestry
115 analysis (SPA), which identifies SNPs with significant gradients in allele frequency (Yang,
116 Novembre, Eskin, & Halperin, 2012). The geographical location (longitude and latitude)
117 information for each tree was supplied as the “--location-input”. SNPs with SPA scores above
118 the 99.9% percentile were considered as outliers. Another outlier detection method was

119 implemented by the OutFLANK software (Whitlock & Lotterhos, 2015). It infers the F_{ST}
120 distribution for a large set of loci and identifies the loci that may contribute to a significant
121 local differentiation and potential adaptation. A Q -value of 0.05 was applied to detect outliers.
122 Following the program recommendation, 1,323,910 SNPs with a minor allele frequency
123 (MAF) ≥ 0.05 were used for the SPA and OutFLANK analyses.

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

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

139 selection signatures was based on univariate models and the threshold for screening
140 significant models was set to 1%.

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

152 The annotation for genes that contain identified SNPs was obtained from loblolly pine
153 gene annotation files available on
154 <https://treegenesdb.org/FTP/Genomes/Pita/v1.01/annotation/> (Wegrzyn et al., 2014). The
155 regulatory sequences including promoters, enhancers and silencers have not yet been
156 identified. SNPs within 5000 bp downstream or upstream of a gene were considered to be
157 within a putative regulatory sequence of the gene. If a SNP is located in a region without
158 annotation, the flanking sequence 700 bp upstream and downstream of the SNP was used as a

159 query to do a blastx search against the entire National Center for Biotechnology Information
160 (NCBI) nonredundant (nr) protein database (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>). The
161 VCFtools software (Danecek et al., 2011) was used to calculate the MAF.

162 **3 | RESULTS**

163 **3.1 | Climate variation in the loblolly pine natural range**

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

178 explaining 62.6% of the variation of the climate variables. The second PC was mainly
179 correlated with precipitation variables, explaining 21.4% of the variation of the climate
180 variables.

181 **3.2 | SNPs associated with climate and geographic variables**

182 We identified 503 associations, including 49 climate/geographic variables and 293 SNPs
183 (Table S2). Among them, 297 associations involved temperature variables, 174 - precipitation
184 variables, 21 - elevation, and 11 - latitude. The MAF of the identified SNPs were between
185 0.01 and 0.5 with a median of 0.02. Among the 293 SNPs, 199 were in 195 annotated genes.
186 Specifically, 3 SNPs (2%) were in 3' regulatory sequences (3' RS), 9 (4%) in 5' RS, 118
187 (59%) in coding sequences (CDS), 59 (29%) in introns, 5 (3%) in 5' untranslated regions (5'
188 UTR), and 5 (3%) in 3' UTR. The remaining SNPs were in unclassified or intergenic regions.
189 Most identified SNPs were associated with multiple variables. For example, the SNP
190 tsc scaffold3881_229913 was associated with latitude, 3 precipitation variables, and 25
191 temperature variables. This SNP resides in the CDS of a gene encoding EARLY
192 FLOWERING 3-like protein, which is a circadian clock protein playing key roles in
193 adaptation of plants to diurnal environmental conditions.

194 **3.3 | Outlier SNPs**

195 We found that 1,324 SNPs showed large gradients in allele frequency based on the SPA
196 analysis (Table S3). Among them, 1,099 SNPs resided in 381 annotated genes. Specifically,

197 43 SNPs (4%) resided in 3' RS, 68 (6%) in 5' RS, 548 (50%) in CDS, 380 (35%) in introns,
198 14 (1%) in 5' UTR, and 46 (4%) in 3'UTR. The other SNPs resided in unclassified or
199 intergenic regions. The annotated genes PITA_000021128 and PITA_000021125 contained
200 the most outlier SNPs, 38 and 27, respectively. These two genes encode the ent-copalyl
201 diphosphate synthase, and the abietadienol/abietadienal oxidase-like protein, respectively.
202 Both genes participate in terpenoid synthesis and contribute to conifer defense against
203 herbivores and pathogens.

204 We also identified 242 SNP outliers using the OutFLANK software (Table S4). Among
205 them, 189 SNPs resided in 128 annotated genes. Specifically, 8 SNPs (4%) resided in 3' RS,
206 11 (6%) in 5' RS, 120 (64%) in CDS, 44 (23%) in introns, 2 (1%) in 5' UTR, and 4 (2%) in
207 3'UTR. The remaining SNPs resided in unclassified or intergenic regions. The annotated
208 genes PITA_000091177, PITA_000064023, and PITA_000040532 contained the most outlier
209 SNPs. These three genes encode a LRR receptor-like serine/threonine-protein kinase, a bHLH
210 transcription factor, and a protein of unknown function.

211 We found 33 loci identified by both SPA and OutFLANK software (Table S5). The MAFs
212 of these 33 loci ranged between 0.06 and 0.47 with a median of 0.21. These 33 loci resided in
213 12 annotated genes encoding proteins that include the leucine-rich repeat receptor-like
214 serine/threonine-protein kinase, the bHLH transcription factor, oxidoreductase, and an
215 EARLY FLOWERING 3-like protein.

216 **3.4 | Multivariate analyses of the identified SNP outliers**

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

228 **3.5 | Outlier SNPs associated with climate and geographic variables**

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

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

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

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

253 **3.6 | SNPs associated with both climate/geographic variables and adaptive** 254 **phenotypic traits**

255 We identified five envSNPs associated with both height and diameter, 10 with height only,
256 114 with 27 metabolite levels, and 242 with expression levels of 47 genes (Tables 2, S8 and
257 S9). For example, 54 envSNPs associated with arachidic acid levels, and more than 60
258 envSNPs associated with the expression levels of *ANR* and *NCED* genes.

259 We combined genomic, phenotypic and climate/geographic data to analyze adaptive
260 genetic variation. For example, we found the envSNP scaffold10517.2_56785 (identified by
261 both association and outlier detection methods) correlated with expression levels of the *ANR*
262 and *NCED* genes. The expression levels of these two genes also correlated with precipitation
263 of May (Figure 5a). The *ANR* gene encodes an anthocyanidin reductase, which is important
264 for the biosynthesis of condensed tannins (Xie, Sharma, Paiva, Ferreira, & Dixon, 2003). The
265 *NCED* gene encodes a 9-*cis* epoxycarotenoid dioxygenase, which prepares precursors for
266 synthesis of abscisic acid (ABA) (Tan et al., 2003). ABA is a key regulator of seed
267 development, root growth, stomatal aperture and plant responses to water stress. The envSNP
268 scaffold10517.2_56785 resided in a gene encoding an abietadienol/abietadienal oxidase-like
269 protein, which is a multifunctional and multisubstrate cytochrome P450 monooxygenase that
270 contributes to conifer defense by generating an enormous structural diversity of plant
271 terpenoid secondary metabolites (Ro, Arimura, Lau, Piers, & Bohlmann, 2005). Individuals
272 with the AA genotype tended to have low expression of the *ANR* gene and high expression of
273 the *NCED* gene (Figure 5b). They were common in counties with low precipitation in May.
274 On the contrary, individuals with the GG genotype had high expression of the *ANR* gene, and

275 low expression of the *NCED* gene. They were common in counties with high precipitation in
276 May. Individuals with the AG genotype were common in counties with medium precipitation
277 in May, and the expression of the *ANR* and *NCED* genes did not differ much from the
278 individuals with the AA genotypes. Precipitation in May positively correlated with the *ANR*
279 gene expression level ($r = 0.4$, $P < 0.0001$) and negatively correlated with the *NCED* gene
280 expression level ($r = -0.2$, $P=0.0005$).

281 **4 | DISCUSSION**

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

288 **4.1 | Comparison of multiple analysis methods**

289 Combining environmental association analyses with outlier detection methods is a desirable
290 way to reduce the rate of false positives and assess the relevance of findings in landscape
291 genomic research (Le Corre & Kremer, 2012; Rellstab et al., 2015), but each method has its
292 strengths and weaknesses. TASSEL exploits the genomic diversity at a very high resolution,
293 hence it is sensitive for detecting associations even for SNPs with low MAFs. In this study,

294 among the 293 envSNPs that demonstrated significant associations with climate and
295 geographic variables detected by TASSEL, 72% had a MAF less than 0.05. Associations
296 could be due to linkage disequilibrium with the functional loci and hence not directly
297 involved in environmental adaptation. The SPA and OutFLANK software detect SNPs under
298 strong selection. To apply these two methods, loci with low MAFs (< 0.05) were removed
299 due to a probable high sampling variance, which may negatively affect the power of models.
300 This is especially critical for OutFLANK, because the distribution of F_{ST} for loci with low
301 MAFs is very different from that for loci with more equal allele frequencies (Whitlock &
302 Lotterhos, 2015). The MAFs of SNPs detected by SPA ranged from 0.06 to 0.5 with a median
303 of 0.36. The MAFs of SNPs detected by OutFLANK ranged from 0.05 to 0.47 with a median
304 of 0.07. Since most adaptation related traits are polygenic with small allele frequency changes
305 at many loci (Le Corre & Kremer, 2012; Mackay, Stone, & Ayroles, 2009), SPA and
306 OutFLANK would miss those loci under weak selection. Additionally, SPA and OutFLANK
307 cannot identify the specific factors that drive selection. To further determine the selective
308 factors, the Samβada software was applied to associate climate and geographic variables with
309 SNP outliers while taking into account spatial autocorrelation. The Bonferroni correction
310 implemented in the current Samβada software may be overly-conservative and may result in
311 overlooking potentially adaptive loci (Stucki et al., 2017). We applied the multivariate
312 approach RDA to examine the relationship between climate/geographic variables and genetic
313 variation of the outlier SNPs. We identified precipitation factors as the important drivers for

314 local adaption. However, the joint effect of climate and geography due to collinearity
315 comprises 49% of the SNP outlier variance. The strong pattern of collinearity could skew the
316 results (Rellstab et al., 2015).

317 The overlap rate among the SNPs detected by different software was relatively low.
318 Among the 1324 and 242 SNP outliers detected by SPA and OutFLANK, respectively, only
319 33 SNPs were the same. Among the 293 and 323 envSNPs identified by TASSEL and
320 Samβada, respectively, only 5 envSNPs were the same. Different assumptions and models
321 applied in different software cause the relatively low numbers of consensus envSNPs. The
322 low consistency across different genome scan methods was also reported previously (de
323 Villemereuil, Frichot, Bazin, François, & Gaggiotti, 2014). There is no single widely accepted
324 statistical approach (Rellstab et al., 2015). Integrating multiple methods and compiling all
325 possible results can provide more reliable information for downstream analyses. Follow-ups
326 are needed to validate the detected adaptive loci and genes using independent populations,
327 knockout mutants, common garden, and reciprocal transplant experiments (Rellstab et al.,
328 2015).

329 **4.2 | Evidence of selection by environment**

330 The identified SNP-environment associations helped us recognize the climate and geography
331 variables that have shaped the genetic variation. We found that longitude, maximum
332 temperature of the warm months and monthly precipitation were variables associated with the

333 most envSNPs (Figure 4). They acted as selective factors driving loblolly pine local
334 adaptation. Although the seed transfer guidelines advised the yearly average minimum
335 temperature as the most important climate variable for southern pines (Schmidtling, 2003),
336 the current study highlights the importance and complexity of maximum temperature of the
337 warm months and monthly precipitation in shaping the genetic variation underlying loblolly
338 pine adaptability. A significant increase in the number of consecutive days exceeding 35°C (a
339 metric used as a measure of heat waves) and a decline in the net water supply availability are
340 expected over the next decades, particularly in the western part of the loblolly pine range
341 (Kunkel et al., 2013; Sun et al., 2013). In a rapid climate change scenario, if adaptation of
342 loblolly pine cannot match the increased heat and drought conditions, the productivity and
343 thus the economic and ecological profits will be greatly damaged. Selecting and planting
344 genotypes adapted to the changing climate may reduce losses in loblolly pine plantations.

345 The identified candidate genes directly or indirectly related to abiotic or biotic stress
346 response, including four functional groups: 1) terpenoid synthesis, 2) pathogen and disease
347 defense, 3) transcription factors, and 4) abiotic stress response (Tables 1 and S7). For
348 example, genes encoding the myrcene synthase and cytochrome P450 are in the terpenoid
349 biosynthesis pathway. Terpenes offer chemical defense against herbivores and pathogens in
350 conifers. The gene encoding a LRR receptor-like serine/threonine-protein kinase is related to
351 pathogen and disease resistance. The transcription factors bHLH and MADS-box regulate
352 downstream defensive and developmental reactions. Other genes are related to responses to

353 abiotic stresses, including stresses from UV, salt, drought, nitrogen, cold, heat, oxidation and
354 wounding. These stress response genes contribute to the genetic structure of loblolly pine
355 adaptability, conferring mitigation and adaptation potential in diverse environments. Five
356 genes related to loblolly pine adaptability and detected in the current study were also reported
357 earlier in Eckert et al. (2010a). These consistently detected genes encode the MATE efflux
358 family protein, a methyltransferase, a translation initiation factor, an ubiquitin, and an auxin
359 responsive protein. They are associated with multiple climate and geographic variables
360 including longitude, monthly precipitation and average maximum monthly temperature. For
361 example, the gene encoding the MATE efflux family protein was previously identified to
362 correlate with aridity (Eckert et al., 2010a). In the current study, this gene was found to be
363 associated with average maximum temperature in February and March, precipitation in
364 January, February, April, June, November and December, mean temperature of the driest
365 quarter (BIO9), annual precipitation (BIO12) and precipitation of the coldest quarter (BIO19).
366 The MATE efflux family proteins play important roles in a wide range of biological
367 processes, such as transporting secondary metabolites, regulating disease resistance and
368 detoxifying toxic compounds (Liu, Li, Wang, Gai, & Li, 2016). These consistently detected
369 genes are strong candidates underlying loblolly pine adaptability.

370 Combining environmental association analyses with dissection of phenotypic traits can
371 greatly improve our understanding of the genetic basis of local adaptation. Talbot et al. (2017)
372 reported that loci with local adaptation signatures in loblolly pine were also linked to gene

373 expression traits for lignin development and whole-plant traits. In our study, more
374 associations between loci with local adaption signatures and adaptive phenotypic traits were
375 detected due to the application of 2.8 million SNPs. The loci with local adaption signatures
376 correlated with height, diameter, metabolite levels, and expression of genes. These results
377 indicate that genes underlying adaptive phenotypic traits are likely involved in adaptability to
378 the environment. These candidate genes need to be further tested in validation populations
379 located in different environments.

380 **5 | CONCLUSION**

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

392 elucidation of the genetic architecture of environmental adaptation in loblolly pine. The
393 knowledge can be applied in breeding loblolly pine trees adapted to the future local
394 environment.

395 **ACKNOWLEDGEMENTS**

396 We thank the Allele Discovery of Economic Pine Traits 2 (ADEPT 2) project (National
397 Science Foundation Grant DBI-0501763) for developing the population, the PineRefSeq
398 Project (USDA National Institute of Food and Agriculture, Award #2011-67009-30030) for
399 providing the draft loblolly pine reference sequences, exon annotation and bioinformatics
400 assistance, and the Texas A&M Institute for Genome Sciences and Society (TIGSS) for
401 providing computational resources and system administration support for the TIGSS HPC
402 Cluster. This study was funded by the Pine Integrated Network: Education, Mitigation, and
403 Adaptation Project (PINEMAP), a Coordinated Agricultural Project funded by the USDA
404 National Institute of Food and Agriculture, Award #2011-68002-30185.

405 **DATA ACCESSIBILITY**

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

410 **AUTHOR CONTRIBUTIONS**

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

415 **DISCLOSURE DECLARATION**

416 The authors declare no competing interest.

417 **ORCID**

418 *Konstantin V. Krutovsky* <http://orcid.org/0000-0002-8819-7084>

419 *Mengmeng Lu* <https://orcid.org/0000-0001-5023-3759>

420 **REFERENCES**

- 421 Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., ...
422 Cobb, N. (2010). A global overview of drought and heat-induced tree mortality
423 reveals emerging climate change risks for forests. *Forest Ecology and Management*,
424 259, 660-684. doi: 10.1016/j.foreco.2009.09.001
- 425 Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007).
426 TASSEL: software for association mapping of complex traits in diverse samples.
427 *Bioinformatics*, 23, 2633-2635. doi: 10.1093/bioinformatics/btm308
- 428 Chhatre, V. E., Byram, T. D., Neale, D. B., Wegrzyn, J. L., & Krutovsky, K. V. (2013). Genetic
429 structure and association mapping of adaptive and selective traits in the east Texas
430 loblolly pine (*Pinus taeda* L.) breeding populations. *Tree Genetics & Genomes*, 9,
431 1161-1178. doi: 10.1007/s11295-013-0624-x
- 432 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ...
433 1000_Genomes_Project_Analysis_Group (2011). The variant call format and

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

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

438 Eckert, A. J., van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González -
439 Martínez, S. C., & Neale, D. B. (2010a). Patterns of population structure and
440 environmental associations to aridity across the range of loblolly pine (*Pinus taeda*
441 L., Pinaceae). *Genetics*, 185, 969-982. doi: 10.1534/genetics.110.115543

442 Eckert, A. J., Bower, A. D., González - Martínez, S. C., Wegrzyn, J. L., Coop, G., & Neale, D. B.
443 (2010b). Back to nature: ecological genomics of loblolly pine (*Pinus taeda*,
444 Pinaceae). *Molecular Ecology*, 19, 3789-3805. doi: 10.1111/j.1365-
445 294X.2010.04698.x

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

449 Kunkel, K. E., Stevens, L. E., Stevens, S. E., Sun, L., Janssen, E., Wuebbles, D., ... Dobson, J. G.
450 (2013). *NOAA technical report NESDIS 142-2. Regional climate trends and scenarios*
451 *for the U.S. national climate assessment. Part 2. Climate of the Southeast U.S.*
452 Washington, DC: National Environmental Satellite, Data, and Information Service.

453 Le Corre, V., & Kremer, A. (2012). The genetic differentiation at quantitative trait loci under
454 local adaptation. *Molecular Ecology*, 21, 1548-1566. doi: 10.1111/j.1365-
455 294X.2012.05479.x

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

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

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

466 Lu, M., Seeve, C. M., Loopstra, C. A., & Krutovsky, K. V. (2018). Exploring the genetic basis
467 of gene transcript abundance and metabolite levels in loblolly pine (*Pinus taeda* L.)
468 using association mapping and network construction. *BMC genetics*, 19, 100. doi:
469 10.1186/s12863-018-0687-7

470 Mackay, T. F., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits:
471 challenges and prospects. *Nature Reviews Genetics*, 10, 565-577. doi:
472 10.1038/nrg2612

473 Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H.
474 (2017). *vegan: Community Ecology Package.R package version 2.4-2.*

475 <https://CRAN.R-project.org/package=vegan>
476 R_Core_Team (2017). R: A language and environment for statistical computing.
477 <https://www.R-project.org/>
478 Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical
479 guide to environmental association analysis in landscape genomics. *Molecular*
480 *Ecology*, 24, 4348-4370. doi: 10.1111/mec.13322
481 Ro, D. K., Arimura, G. I., Lau, S. Y., Piers, E., & Bohlmann, J. (2005). Loblolly pine
482 abietadienol/abietadienal oxidase *PtAO* (CYP720B1) is a multifunctional,
483 multisubstrate cytochrome P450 monooxygenase. *Proceedings of the National*
484 *Academy of Sciences of the United States of America*, 102, 8060-8065. doi:
485 10.1073/pnas.0500825102
486 Schmidting, R. C. (1988). Racial variation in self-thinning trajectories in loblolly pines. In
487 *Proceedings of the IUFRO Conference*, pp. 611-618.
488 Schmidting, R. C., & Froelich, R. C. (1993). Thirty-seven year performance of loblolly pine
489 seed sources in eastern Maryland. *Forest Science*, 39, 706-721.
490 Schmidting, R. C. (2001). *General technical report SRS-44. Southern pine seed sources.*
491 Asheville, NC: USDA Forest Service, Southern Research Station.
492 Schmidting, R. C. (2003). Determining seed transfer guidelines for southern pines. In L. E.
493 Riley, R. K. Dumroese & T. D. Landis, technical coordinators, *National proceedings:*
494 *Forest and conservation nursery associations-2002. Proceedings RMRS-P-28.* (pp. 8-
495 11). Ogden, UT: USDA Forest Service, Rocky Mountain Research Station.
496 Sork, V. L., Aitken, S. N., Dyer, R. J., Eckert, A. J., Legendre, P., & Neale, D. B. (2013). Putting
497 the landscape into the genomics of trees: approaches for understanding local
498 adaptation and population responses to changing climate. *Tree Genetics & Genomes*,
499 9, 901-911. doi: 10.1007/s11295-013-0596-x
500 Stucki, S., Orozco - terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., ... Joost, S.
501 (2017). High performance computation of landscape genomic models including
502 local indicators of spatial association. *Molecular Ecology Resources*, 17, 1072-1089.
503 doi: 10.1111/1755-0998.12629
504 Sun, G., Caldwell, P. V., McNulty, S. G., Georgakakos, A. P., Arumugam, S., Cruise, J., ... Marion,
505 D. A. (2013). *NCA Southeast technical report. Impacts of climate change and*
506 *variability on water resources in the Southeast USA.* Asheville, NC: USDA Forest
507 Service, Southern Research Station.
508 Talbot, B., Chen, T.-W., Zimmerman, S., Joost, S., Eckert, A. J., Crow, T. M., ... Manel, S. (2017).
509 Combining genotype, phenotype, and environment to infer potential candidate
510 genes. *Journal of Heredity*, 108, 207-216. doi: 10.1093/jhered/esw077
511 Tan, B. C., Joseph, L. M., Deng, W. T., Liu, L., Li, Q. B., Cline, K., & McCarty, D. R. (2003).
512 Molecular characterization of the *Arabidopsis* 9-*cis* epoxycarotenoid dioxygenase
513 gene family. *The Plant Journal*, 35, 44-56. doi: 10.1046/j.1365-313X.2003.01786.x
514 Wear, D. N., Huggett, R., Li, R., Perryman, B., & Liu, S. (2013). *Gen.Tech.Rep.SRS-GTR-170.*
515 *Forecasts of forest conditions in regions of the United States under future scenarios:*

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

518 Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L. S., Loopstra, C. A., Vasquez-Gross, H. A., ...
519 Neale, D. B. (2014). Unique features of the loblolly pine (*Pinus taeda* L.)
520 megagenome revealed through sequence annotation. *Genetics*, *196*, 891-909. doi:
521 10.1534/genetics.113.159996

522 Wells, O. O. (1985). Use of Livingston Parish, Louisiana loblolly pine by forest products
523 industries in the Southeast. *Southern Journal of Applied Forestry*, *9*, 180-185. doi:
524 10.1093/sjaf/9.3.180

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

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

531 Yang, W.-Y., Novembre, J., Eskin, E., & Halperin, E. (2012). A model-based approach for
532 analysis of spatial structure in genetic data. *Nature Genetics*, *44*, 725-731. doi:
533 10.1038/ng.2285

534

535 **SUPPORTING INFORMATION**

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