Exploring the genetic basis of gene transcript abundance and metabolite level in loblolly pine (*Pinus taeda* L.) using association mapping and network construction

Mengmeng Lu[1, 2, 3], Candace M. Seeve[4], Konstantin V. Krutovsky[1, 2, 5, 6, 7], Carol A. Loopstra[1, 2, 8]

[1]Department of Ecosystem Science and Management, Texas A&M University, 2138 TAMU, College Station, TX 77843-2138, USA; [2]Molecular and Environmental Plant Sciences Program, Texas A&M University, 2474 TAMU, College Station, TX 77843-2474, USA; [3]Present address: Department of Biological Sciences, University of Calgary, 507 Campus Drive NW, Calgary, AB T2N 4S8, Canada; [4]USDA-ARS Midwest Area, Columbia, MO 65211, USA; [5]Department of Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Büsgenweg 2, Göttingen 37077, Germany; [6] Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina Str. 3, Moscow 119333, Russia; [7]Genome Research and Education Center, Siberian Federal University, 50a/2 Akademgorodok, Krasnoyarsk 660036, Russia; [8]Corresponding author, email: c-loopstra@tamu.edu, phone: 979-862-2200

Keywords: gene expression, metabolism, epistasis, stress response, wood development, SNP

Abstract

Gene transcripts and metabolites are important regulatory checkpoints between genetic variation and complex biological processes such as wood development and drought response in conifers. Loblolly pine (*Pinus taeda* L.) is one of the most commonly planted forest tree species in the southern U.S. In this study, we tested for associations between 2.8 million exome-derived SNPs and the transcript abundance of 110 wood development genes, 88 disease or drought related genes as well as levels of 82 known metabolites. We identified 1841 SNPs associated with 191 gene expression phenotypes and 524 SNPs associated with 53 metabolite level phenotypes. The identified SNPs reside in genes with a wide variety of functions. We further integrated the identified SNPs and their associated expressed genes and metabolites into networks. We described the SNP-SNP interactions that significantly impacted the gene transcript abundance and metabolite level in the networks. The key loci and genes in the wood development and drought response networks were identified and analyzed. This work provides candidate genes for research on the genetic basis of gene expression and metabolism linked to wood development and drought response in loblolly pine, and highlights the efficiency of using association-mapping-based networks to discover candidate genes with important roles in complex biological processes.

Keywords: gene expression, metabolism, epistasis, stress response, wood development, SNP

Introduction

Understanding the genetic basis of complex traits in the important forest tree species, loblolly pine (*Pinus taeda* L.), can contribute to the improvement of its growth and quality. The majority of previous genetic studies have focused on the dissection of adaptive or commercially important traits like growth, wood properties, or drought tolerance (Neale and Savolainen 2004; González-Martínez et al. 2007; Cumbie et al. 2011; Westbrook et al. 2013), while only a few studies have sought to characterize phenotypes in depth by surveying the levels of transcripts and metabolites associated with such traits of interest. Palle et al. (2011) analyzed expression of genes involved in loblolly pine wood development and reported key regulatory genes. A total of 33 wood development gene expression phenotypes were associated with 80 single nucleotide polymorphisms (SNPs). Seeve (2010) detected the expression of 88 genes related to disease or drought responses in loblolly pine and found that 27 expression phenotypes were associated with 94 SNPs. Eckert et al. (2012) detected multiple SNP-metabolite associations in loblolly pine.

Gene transcript abundance and metabolite levels are complex intermediate phenotypes that link genetic variations to whole-plant phenotypes. Each is regulated by genetic and environmental cues, and perturbations in these intermediate phenotypes may be manifested as changes in higher-order traits (Schadt et al. 2008). Thus, studies linking gene expression or metabolite phenotypes to genetic variations may enhance our understanding of the molecular mechanisms that underlie broader

whole-plant phenotypes. For example, Bossu et al. (2016) found secondary metabolites influence wood properties. Obata et al. (2015) demonstrated that metabolite levels in maize respond to stress conditions and can be used to predict the grain yield under drought. Furthermore, integrating SNPs and their associated gene expression and metabolite level phenotypes into networks aids in connecting the two phenotypes, and in identifying key genes in regulatory networks that contribute to adaptive traits (Wentzell et al. 2007; Burkhardt et al. 2015).

To gain insights into the regulatory mechanism underlying wood development and disease and drought responses, we tested for associations between 2.8 million SNPs derived from exome target sequencing and gene transcript abundance and metabolite levels. The expression data includes 110 wood development genes and 88 disease or drought related genes. The metabolite data includes 82 metabolites with known names. We constructed networks to analyze the loci associated with multiple phenotypes. Since epistatic interaction between loci is another factor that may further influence phenotypes in loblolly pine (Lu et al. 2017), the SNP-SNP interactions were also detected among the identified loci. The identified genes are valuable resources to study the genetic basis of gene expression and metabolite level phenotypes linked to complex biological processes in loblolly pine.

Materials and methods

Plant material and genotypic data

The loblolly pine population used in this study was originally established for the Allele Discovery of Economic Pine Traits 2 (ADEPT2) project and included trees with parents from a wide range across the southeastern U.S. (Eckert et al. 2010a; Cumbie et al. 2011). Genotypic data were obtained for 375 trees in this 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 were described in Lu et al. (2016). 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) $\geq$ 0.01. A total of 2,822,609 SNPs were retained, and a total of 94,478 haplotype blocks were detected for this population (Lu et al. 2017). Additionally, 23 simple sequence repeat (SSR) markers have been used to genotype ADEPT2 trees (Eckert et al. 2010a). SSR genotype data were used for estimating covariates to adjust for the selectively neutral population structure.

Phenotypic data

Abundance of functional gene transcripts and levels of metabolites were analyzed in this study. Relative transcript abundance was measured using reverse transcription quantitative polymerase chain reaction (RT-qPCR). Palle et al. (2011) analyzed the

109 expression of 111 genes with probable roles in xylem/wood development in woody
110 tissue collected from 475 trees. Seeve (2010) detected the expression of 88 disease or
111 drought responsive genes in woody tissue collected from 354 trees. However, only
112 278 trees with gene expression data were genotyped for this study. Therefore, 278
113 trees were used for association tests with expression data for 199 genes. The gene
114 expression phenotypes from the two data sets were organized into seven functional
115 groups based on the biological processes which they were involved: genes related to
116 reactive oxygen species (ROS) biosynthesis and signaling, terpenoid biosynthesis,
117 programmed cell death (PCD), phenylpropanoid pathway, wood-related,
118 disease-related, and drought-related genes. The genes in each group were further
119 assigned to sub-groups (see Table S1 available as Supplementary Data at *Tree*
120 *Physiology* Online). Metabolite data were obtained from the study of Eckert et al.
121 (2012). They measured the concentration of 292 metabolites in woody tissue of
122 ADEPT2 trees. In this study, we only used data of the 82 metabolites with known
123 names. Only 212 of the trees with metabolite data were genotyped for this study.
124 Therefore, 212 trees were used for association tests with concentration data for 82
125 metabolites.
126
127 ***Association analyses***
128 Association analyses for the individual SNPs and phenotypes were conducted using
129 TASSEL 5.0 (Bradbury et al. 2007). The SSR genotype data were used for estimating

130 covariates to adjust for the selectively neutral population structure. The SSR
131 genotypes were available for 195 of the trees used for the gene expression analysis
132 and 196 of the trees used for the metabolite concentration analysis. We used this
133 group of trees (named as the *str* population) for a population structure analysis.
134 Population structure within this group was mainly due to the Mississippi River
135 discontinuity (Lu et al. 2016). We named the trees from east of the Mississippi River,
136 namely 223 trees used for gene expression analysis and 184 trees used for metabolite
137 concentration analysis, as the *east* population. Therefore, three populations: *total* (N =
138 278), *east* (N = 223) and *str* (N = 195) populations, were used to perform association
139 analyses for the gene expression data. Three populations, *total* (N = 212), *east* (N =
140 184) and *str* (N = 196), were used to perform association analyses for the metabolite
141 concentration data. For the *total* and *east* populations, the simple general linear model
142 (GLM) method (*S* model) and the mixed linear model (MLM) method incorporating a
143 kinship matrix (*K* model) were applied. For the *str* population, in addition to the *S* and
144 *K* models, the GLM incorporating the covariate to adjust for population structure (*Q*
145 model) and the MLM incorporating both the kinship matrix and population structure
146 covariate (*QK* model) were applied. The population structure covariate was estimated
147 using the software STRUCTURE (Pritchard et al. 2000; Hubisz et al. 2009) and 23
148 SSR markers. A kinship matrix for each population was estimated by TASSEL 5.0
149 (Bradbury et al. 2007) using the 2.8 million SNP markers. The kinship relatedness is
150 low in this population with an average range between -0.03 and 0.10 (excluding the

151 self-relatedness). Quantile-quantile plots were generated for observed against
152 expected $-\log_{10}P$ to examine the model fitness, where observed *P*-values were
153 obtained from association mapping and expected *P*-values from the assumption that
154 no association occurred between marker and trait. Significance of associations
155 between loci and traits were determined by the *P*-values. A corrected Bonferroni
156 threshold 0.05/94,478=5.29E-7, where 94,478 was the estimated number of haplotype
157 blocks, was applied to screen for significant loci. The squared correlation coefficient
158 (R^2) between genotypes on the same scaffold was used as an LD measure and
159 calculated using the "geno-r2" function in the VCFtools software (Danecek et al.
160 2011). The triangular heatmaps were produced using the R package "LDheatmap"
161 (Shin et al. 2006; R Core Team 2017).
162
163 ***Annotation of genes that contained SNPs associated with traits***
164 The VCFtools software (Danecek et al. 2011) was used to calculate the minor allele
165 frequencies (MAFs) and perform Hardy-Weinberg Equilibrium (HWE) tests for the
166 identified SNPs. Annotation of the genes containing the identified SNPs was obtained
167 from loblolly pine Gene Annotation v3.0
168 (https://treegenesdb.org/FTP/Genomes/Pita/v1.01/annotation) (Wegrzyn et al. 2014).
169 Very few regulatory sequences such as promoters, enhancers and silencers have been
170 identified in the loblolly pine reference genome. SNPs within 5000 bp downstream or
171 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 1500 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) non-redundant (nr) protein database (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The NCBI GI numbers of candidate genes were uploaded to the "Gene List Analysis" tool in the PANTHER Classification System (http://www.pantherdb.org) (Mi et al. 2013; Mi et al. 2016). The genes were mapped to the PANTHER databases and analyzed for their classification according to their molecular functions and protein classes.

Network plots and SNP-SNP interaction analyses

To visualize the relationships between SNPs and their associated phenotypes, R package "igraph" was used to plot the networks (Csardi and Nepusz 2006; R Core Team 2017). Blue, yellow and pink nodes represent SNPs, gene expression phenotypes and metabolite level phenotypes, respectively. Red and gray edges represent the significant SNP-metabolite-level and SNP-gene-expression associations. In addition, for the SNPs in the networks, the epistatic SNP-SNP interaction test was implemented using PLINK 1.9 (Purcell et al. 2007). The Bonferroni correction was applied to screen for significant SNP-SNP interactions. In the networks, purple edges represent the significant SNP-SNP interactions.

193 ## Results

194 ***Significant associations between SNPs and phenotypes***

195 Association analyses of 2.8 million SNPs with 199 gene expression phenotypes and
196 82 metabolite level phenotypes were conducted. After summarizing the results from *S*,
197 *K*, *Q* and *QK* models, a total of 2,562 associations between 1,841 SNPs and 191 gene
198 expression phenotypes and 524 associations between 524 SNPs and 53 metabolite
199 concentration phenotypes were identified (see Tables S2 & S3 available as
200 Supplementary Data at *Tree Physiology* Online). A total of 40 % and 23 % of the
201 SNPs associated with gene expression and metabolite concentration phenotypes,
202 respectively, had a MAF ≥ 0.05. The MAFs of other SNPs were between 0.01 and
203 0.05. There were 9 % of the SNPs associated with gene expression and 6 % of the
204 SNPs associated with metabolite concentrations that departed from HWE. Among the
205 2562 gene expression associations, 1195 (47 %) were related to expression of wood
206 development genes, 661 (26 %) to drought-related genes, 232 (9 %) to terpenoid
207 biosynthesis genes, 162 (6 %) to PCD genes, 110 (4 %) to ROS genes, 104 (4 %) to
208 phenylpropanoid pathway genes and 98 (4 %) to disease related genes. Expression of
209 the *CYPB* gene (involved in terpenoid biosynthesis) was associated with the largest
210 number of SNPs (181 SNPs). It was followed by the genes encoding *ATAF-1* [a
211 drought-responsive transcription factor (TF), 138 SNPs], *RAP2.1* (a
212 drought-responsive TF, 133 SNPs), *CS-5828* (cellulose synthase-like, 128 SNPs),
213 *CslA1* (cell wall- related, 117 SNPs), *PtEMB4* (a late embryogenesis abundant protein,

114 SNPs), *αtub1* (α-tubulin, 105 SNPs), *ANR* (involved in phenylpropanoid pathway, 76 SNPs), *PtMLO2* (involved in PCD, 75 SNPs), *NCED* (related to drought signaling, 73 SNPs), *PtMLO1* (involved in PCD, 74 SNPs), *CesA2* (a cellulose and callose synthase, 66 SNPs), *RP-L2* (a wood development protein, 62 SNPs), and *CaS3* (a cellulose and callose synthase, 57 SNPs). For levels of the metabolites glucose and melezitose, each were associated with 30 SNPs. They were followed by 3,4-dihydroxybenzoic acid (24 SNPs), glycerol-3-galactoside (22 SNPs), glycine (21 SNPs), and raffinose (21 SNPs). Complete lists of the identified SNPs and their associated phenotypes were presented in Tables S4 & S5 (available as Supplementary Data at *Tree Physiology* Online).

The SNP-trait r^2 values in the association outputs represent the proportion of phenotypic variation that is explained by the corresponding markers. The median of r^2 values was 0.15 for both gene expression and metabolite level associations. However, the r^2 values of gene expression associations had a wide range, from 0.09 to 0.85, while the r^2 values of metabolite level associations ranged from 0.11 to 0.22 (see Figure S1 available as Supplementary Data at *Tree Physiology* Online). We examined the 323 gene expression associations with high r^2 values (> 0.40). A total of 181 were associated with the *CYPB* gene involved in terpenoid biosynthesis, 133 with the *RAP2.1* gene encoding a drought-responsive TF, 4 with the *PtMLO1* gene involved in PCD, 2 with the *PtGPX3* gene (a peroxidase), 2 with the *CesA2* gene (a cellulose and callose synthase), and 1 with the *CaS3* gene (a cellulose and callose synthase).

In the previous association studies on the ADEPT2 population, nearly 4000 EST-derived SNPs were associated with metabolite level and gene expression phenotypes (Seeve 2010; Eckert et al. 2012; Palle et al. 2013). To cross-reference associated SNPs identified in the current study with associated SNPs in the three prior studies, we mapped the sequences with previously identified SNPs to loblolly pine reference assembly v1.01 (https://treegenesdb.org/FTP/Genomes/Pita/v1.01) using the GMAP software (Wu and Watanabe 2005). We found that the SNPs scaffold596656_40783 and tscaffold2197_12732 discovered in the current study reside also in sequences identified in the prior study. The SNP scaffold596656_40783 was associated with expression of the *CAD1* gene (encoding cinnamyl-alcohol dehydrogenase involved in a lignin biosynthesis). This SNP resides in a gene encoding cystathionine gamma-synthase. The SNP tscaffold2197_12732 was associated with expression of the *CesA2* gene (encoding a cellulose and callose synthase). This SNP resides in a gene encoding E3 ubiquitin-protein ligase. These two associations are consistent with the associations reported by Palle et al. (2013). Other identified SNPs in this study could not be mapped to the sequences identified in prior studies. Nonetheless, SNPs and genes identified in the current study provide valuable clues to understand the genetic basis of gene transcript abundance and metabolite level in loblolly pine.

Annotation of the genes containing identified SNPs

We obtained the annotation for the genes containing the identified SNPs from loblolly

256 pine Gene Annotation v3.0 or blastx alignment. The SNPs that were associated with
257 gene expression phenotypes reside in 1635 different annotated genes. Of this total, 57 %
258 reside in coding sequences (CDS), 2 % in 5' untranslated sequences (5'UTR), 3 % in 3'
259 untranslated sequences (3'UTR), 23 % in introns, 7 % in putative 3' regulatory
260 sequences (P3'RS) and 8 % in putative 5' regulatory sequences (P5'RS). The SNPs that
261 were associated with metabolite level phenotypes reside in 374 different annotated
262 genes. Of these, 58 % reside in CDS, 2 % in 5'UTR, 2 % in 3'UTR, 25 % in introns, 6 %
263 in P3'RS, and 7 % in P5'RS. The SNP-containing genes encode proteins with
264 functions of nucleic acid binding, transporter, oxidoreductase, transferase, hydrolase,
265 receptor, enzyme modulator, ligase, cytoskeletal protein, TF, membrane traffic protein,
266 and signaling molecule chaperone. The major molecular functions of SNP-containing
267 genes include: catalytic activity, DNA binding, transporter activity, receptor activity
268 and structural molecule activity.
269 Among the identified associations, some gene expression phenotypes were
270 associated with a large number of SNPs. For example, expression of the *CYPB* gene,
271 which encodes a terpenoid biosynthesis enzyme cytochrome P450 monooxygenase,
272 was associated with 181 SNPs. The SNPs associated with *CYPB* gene expression
273 mainly reside in the genes involved in secondary metabolites biosynthesis and defense
274 resistance, including genes encoding beta-glucosidase, phosphofructokinase,
275 polygalacturonase, shikimate O-hydroxycinnamoyltransferase-like, cytochrome P450
276 78A3, glucosinolate transporter-2, TIR-NBS-LRR protein, serine/threonine protein

277 kinase, and lipase. The expression phenotypes of genes encoding drought-responsive
278 TFs, *RAP2.1* and *ATAF-1*, were also associated with a large number of SNPs, 133 and
279 138 SNPs, respectively. The associated SNPs mainly reside in drought responsive
280 genes or TF genes that confer drought tolerance to plants including genes encoding
281 cysteine-rich receptor-like protein, glucan endo-1,3-beta-glucosidase, COBRA-like
282 protein, cinnamoyl-CoA reductase, root phototropism protein, putative
283 TIR-NBS-LRR protein, laccase, cellulose synthase, UDP-glucuronyltransferase-like
284 protein, and TFs of ethylene-responsive, bHLH, MADS-box and MYBs. Table 1
285 presents a partial list of the genes containing SNPs associated with gene expression
286 and metabolite level phenotypes. More details are presented in Tables S2 & S3
287 (available as Supplementary Data at *Tree Physiology* Online).
288 TFs regulate gene expression in response to a variety of endogenous and
289 environmental cues. The SNP-containing genes that encode TFs were assigned to
290 plant TF families according to the Plant Transcription Factor Database v4.0
291 (http://planttfdb.cbi.pku.edu.cn/index.php). A total of 12 TF families were associated
292 with gene expression and metabolite level phenotypes (Figure 1). Twelve
293 SNP-containing TF family genes belong to MYB family, associating with expressed
294 genes encoding wood development protein (*1CAB-3A*), cellulose and callose synthase
295 (*CesA)*, cell wall protein (*CslA*), α-tubulin (*αtub1*), lignin biosynthesis enzyme
296 (*TC4H*), drought-responsive TF (*RAP2.1*), phenylpropanoid pathway enzyme (*ANR*)
297 and metabolites 4-hydroxybenzoate, aspartic acid, maltose and melezitose. Details of

298 the TFs annotations, SNPs and their associated phenotypes are listed in Table S6
299 (available as Supplementary Data at *Tree Physiology* Online).
300
301 ***LD among identified SNPs that reside in the same scaffolds***
302 Among the identified SNPs, we found that even though some loci are more than 10
303 kbp apart along the same scaffolds, they were associated with the same gene
304 expression phenotypes with similar r^2 values. To examine whether these loci are in
305 linkage disequilibrium (LD), we calculated their pairwise zygotic LD (squared
306 correlation coefficient R^2) values. From the results, we identified 10 scaffolds
307 containing correlated SNPs. For example, the SNPs tscaffold2867_628232,
308 tscaffold2867_651263, and tscaffold2867_755157 span 128 kbp on tscaffold2867
309 (Figure 2). They all were associated with expression of the *ATAF-1* gene
310 (drought-responsive TF) with $r^2 = 0.31$. High pairwise LD values (> 0.89) were
311 detected between these SNPs.
312 To further inspect for haplotype blocks on these scaffolds, we plotted LD heatmaps
313 for SNPs in these regions with SNPs with high correlation values. Figure 2 illustrates
314 all LD values between SNP pairs around the investigated regions on tscaffold2867.
315 Other LD heatmaps are presented in Figures S2-S10 (available as Supplementary
316 Data at *Tree Physiology* Online). We did not observe long LD blocks along the
317 investigated regions in the LD heatmaps.
318

Network plots

Among the identified SNPs, some are associated with multiple gene expression and metabolite level phenotypes. Plotting these SNPs with the gene expression and metabolite level phenotypes with which they are associated in networks can provide insight into the complex regulatory mechanisms underlying biological processes and help us recognize key genes in the pathways. The network graphs were based on the functional groups we assigned. The effects of SNP-SNP interactions were also demonstrated in the networks.

The wood development and drought response networks (Figures 3 & 4, respectively) contain the largest number of SNPs. In the wood development network (Figure 3), a total of 52 SNPs (each represented as a number in a blue node) are connected to 56 gene expression phenotypes (yellow nodes, grey edges) and 8 metabolite level phenotypes (pink nodes, red edges). In the drought response network (Figure 4), a total of 80 SNPs (each represented as a number in a blue node) are connected to 10 gene expression phenotypes (yellow nodes, grey edges) and 4 metabolite level phenotypes (pink nodes, red edges). In both networks, purple edges denote SNP-SNP interactions that significantly impact the phenotypes. The putative functions of the SNP-containing genes included in these networks were determined from loblolly pine gene annotation (https://treegenesdb.org/FTP/Genomes/Pita/v1.01/annotation) or blastx alignment (Tables 2 & 3 and Tables S7 & S8 available as Supplementary Data at *Tree*

340 *Physiology* Online). SNP #33 in the wood development network (Figure 3) and SNPs
341 #13, #20, #57, #70 and #78 in the drought response network reside in TF genes
342 (Figure 4).
343 Fewer associations between SNPs and gene expression phenotypes belonging to
344 the other functional groups were identified. Therefore, limited connections are shown
345 in the ROS response and disease response networks (see Figure S11, Table S9
346 available as Supplementary Data at *Tree Physiology* Online). No networks could be
347 plotted for gene expression phenotypes related to terpenoid biosynthesis, PCD or the
348 phenylpropanoid pathway.
349 Modules of genes with similar functionality can be recognized from the networks.
350 Gene-module level analysis can help us understand developmental and stress
351 resistance phenotypes in the context of biological network design and system
352 behavior rather than as a product of individual genes (Wang et al. 2008). A large gene
353 module related to wood development can be recognized in Figure 3. It contains 33
354 SNPs, 4 metabolites and 28 expressed genes that encode cellulose and callose
355 synthases, lignin biosynthetic enzymes, wood development enzymes, and tubulins.
356 Figure 4 includes two gene modules linked to drought responsive processes. One
357 module is composed of 24 SNPs, 2 metabolites and 4 expressed genes that encode
358 drought responsive TFs, drought signaling molecules and phenylpropanoid pathway
359 enzymes. The other module contains 52 SNPs and two expressed genes that encode a
360 drought responsive TF and a late embryogenesis abundant protein. These modules

361 supplement current regulatory and biosynthetic pathways for wood development and
362 drought response.
363
364 ## Discussion
365 Genetic variations do not lead to changes in whole-plant traits directly, but instead act
366 on intermediate, molecular phenotypes, which in turn induce changes in higher-order
367 traits (Schadt et al. 2008). Therefore, identification of the genetic variants that
368 associate with intermediate phenotypes and description of molecular networks that
369 genes operate are important to understand the genetic basis underlying complex traits.
370 In this study, we explored the genetic regulation of gene transcript abundance and
371 metabolite level linked to important whole-plant traits, wood development and stress
372 responses by constructing networks comprised of the SNPs and their associated gene
373 expression and metabolite level phenotypes. The SNP-SNP interactions were also
374 described in the networks. These results provide valuable sources to bridge
375 connections between genetic variation, intermediate molecules produced in the
376 biological pathways, and whole-plant traits.
377 We identified 1841 SNPs associated with 191 gene expression phenotypes and 524
378 SNPs associated with 53 metabolite level phenotypes. Compared to a wide range of r^2
379 values of gene expression associations (0.09 to 0.85), we did not find strong
380 association signals for metabolite level associations (0.11 to 0.22), probably because
381 SNP effects for metabolite level are generally low and the genetic basis underlying

382 metabolism involves more complex factors.

383 Among the SNP-gene expression associations, we detected 181 associations with
384 *CYPB* gene expression and 133 associations with *RAP2.1* gene expression that have
385 remarkably high r^2 values, ranging from 0.40 to 0.85. The *CYPB* gene encodes a
386 cytochrome P450 monooxygenase enzyme involved in the synthesis of diverse
387 oleoresin terpenoids important for constitutive and induced defenses against pests and
388 pathogens (Ro et al. 2005), while the *RAP2.1* gene encodes a
389 dehydration-responsive-element binding (DREB) protein type transcriptional
390 repressor. We also detected SNPs in strong associations with other gene expression
391 phenotypes, including the gene encoding abiotic stress responsive TF *ATAF-1*, and the
392 gene encoding phenylpropanoid pathway enzyme *ANR*. High r^2 values indicate that
393 the corresponding markers can explain a large proportion of the variation in
394 expression of these genes, and that the associated SNPs offer potential to discover
395 genes that regulate these biosynthetic pathways and stress responses. The SNPs
396 highly associated with *CYPB* and *RAP2.1* gene expression are found in diverse genes.
397 SNPs associated with *CYPB* gene expression were discovered in genes involved in
398 secondary metabolite biosynthesis and defense pathways, including genes encoding
399 NBS-LRR type disease resistance protein and genes encoding MADS-box TF. SNPs
400 associated with *RAP2.1* gene expression were discovered in drought responsive genes
401 or TF genes that contribute to drought tolerance, such as genes encoding MYB, which
402 plays a great role in controlling responses to biotic and abiotic stresses (Ambawat et

403 al. 2013). Although the effects of genes containing the identified SNPs on the
404 expressed genes need to be confirmed by the evidence from forward genetics
405 experiments, association studies are an efficient method to discover clusters of
406 candidate genes in biosynthetic pathways.
407 The pattern and extent of LD in the genome is important for association mapping
408 studies (Yu et al. 2008). In this study, we detected loci located more than 10kbp apart
409 along the same scaffolds that were associated with the same gene expression
410 phenotypes and had similar r^2 values. This observation raised the possibility that these
411 SNPs are in LD with each other or are even found within LD blocks. Although
412 outcrossing conifer trees are thought to have a rapid decline of LD, the rate of LD
413 decay may vary from gene to gene (Brown et al. 2004; Pavy et al. 2012). Furthermore,
414 if loci associated with the same phenotypes are in LD, it may suggest epistatic
415 interaction between these loci due to natural selection. In the current study, we
416 detected ten scaffolds that contained identified SNPs in strong LD with each other.
417 However, no LD blocks were observed in LD heatmap plots for the regions
418 surrounding the correlated SNPs (Figure 2 & Figures S2-S10 available as
419 Supplementary Data at *Tree Physiology* Online). These results diminish the potential
420 of interaction among investigated loci due to natural selection since large blocks of
421 LD should be maintained, if the interacted loci are under selection (Gabriel et al. 2002;
422 Slatkin 2008). The occasional LD observed here probably rise from mixing of
423 individuals from subpopulations. The population used in this study was comprised of

424 individuals with parents from a wide range across the southeastern U.S. Differences in
425 allele frequencies among subpopulations can create resemblance of LD (Slatkin
426 2008).

427 Gene networks demonstrate the potential interactions among genes and help us
428 prioritize the candidate genes (Li et al. 2015). In the wood development network
429 (Figure 3), SNP#33 resides in a TF *GAMYB* gene. It has been identified as an
430 activator of gibberellin (GA)-regulated genes in plant growth (Woodger et al. 2003).
431 SNP#33 was found to be associated with expressed genes encoding wood
432 development enzyme and lignin biosynthetic enzyme, indicating that the *GAMYB*
433 gene may influence lignin biosynthesis and wood formation through its regulatory
434 interactions with a large number of genes. SNP #17 resides in a gene encoding
435 arabinosyltransferase *ARAD1*. It is responsible for the polymerization of arabinose
436 into the arabinan of arabinogalactan (Belanger et al. 1996). Arabinogalactan protein
437 have been found functional during secondary wall formation in loblolly pine (Zhang
438 et al. 2003). SNP#17 is associated with seven gene expression phenotypes all related
439 to lignin biosynthesis. Lignin biosynthesis can be induced when cell wall is damaged
440 (Denness et al. 2011). The associations between SNP#17 and lignin biosynthesis gene
441 expression phenotypes imply a link between arabinogalactan protein and lignin
442 biosynthesis for cell wall formation. SNP#31 resides in an aspartokinase gene.
443 Aspartokinase is an enzyme that catalyzes the phosphorylation of aspartic acid. Data
444 from bacteria has shown that decreasing aspartokinase activity results in blockage of

445 cell wall growth (Rosenberg et al. 1973). The SNP#31 is associated with multiple
446 lignin biosynthesis and wood development gene expression phenotypes, suggesting
447 aspartokinase-mediated amino acid metabolism is involved in cell wood development
448 and lignin biosynthesis. Laccase provides the oxidative capacity during lignification.
449 The large number of gene family members makes it difficult to study (Piscitelli et al.
450 2010). From the network in the Figure 3, we can identify a series of candidate genes
451 that may function in the laccase synthesis pathway. Lac3 gene expression is
452 associated with SNPs that reside in genes encoding cytochrome, disease resistance
453 protein, calcium dependent protein kinase, LRR receptor-like and aspartokinase. Lac6
454 gene expression is associated with SNPs that reside in genes encoding
455 transmembrance protein, 1-phosphatidylinositol 3-phosphate, arabinosyltransferase
456 and CBL-interacting protein kinase. These associations provide clues to understand
457 the laccase oxidation process.
458 Additive or epistatic interaction between loci is another factor that may further
459 influence phenotypes (Phillips 2008). Lu et al. (2017) reported 11 SNP-SNP
460 interactions in loblolly pine that in some cases, contributed more to the clonal and
461 phenotypic variance of the quantitative traits than the identified additive loci. Thus by
462 integrating SNP-SNP epistatic relationships into the network, we can acquire a more
463 complete understanding of gene interactions. In the wood development network
464 (Figure 3), *RP-L2* (ribosomal protein L2) gene expression is impacted by interactions
465 of multiple SNP-SNP pairs. *RP-L2* together with the 23S RNA are the main

candidates for catalyzing peptide bond formation on the 50S subunit (Diedrich et al. 2000). The SNP-SNP interactions suggest genes encoding dormancy/auxin associated protein, pentatricopeptide repeat-containing protein and histone H2A interact to affect the formation of ribosomal protein. Additionally, interaction between an aspartokinase gene and a disease resistance gene significantly influence *CCoAMT* gene expression, but the mechanism remains unclear.

We also discovered important loci and phenotypes from the drought response network (Figure 4). Four gene expression phenotypes stand in the center of a series of SNP associations. *NCED* is a key enzyme in abscisic acid (ABA) biosynthesis, which is induced by drought stress. *ANR* functions in the phenylpropanoid pathway. Expression of *NCED* and *ANR* genes are widely associated with the same set of SNPs, which reside in genes mainly encoding drought responsive products. This result indicates *ANR* and *NCED* genes play key roles in the drought response pathway. *PtEMB4* is a Late Embryogenesis Abundant protein. *ATAF-1* gene belongs to the *NAC* (No Apical Meristem) family genes, which encode plant-specific TFs involved in diverse biological processes (Wu et al. 2009). We found the expression of *ATAF-1* and *PtEMB4* genes were associated with the same 52 SNPs, which reside in genes encoding proteins such as wall-associated receptor kinase-like, heat stress TF. The above results suggest the *PtEMB4* and *ATAF-1* genes as well as the *NCED* and *ANR* genes may perform redundant functions during drought response processes.

Alternatively, there could be a synergetic mechanism for these genes to function together during drought response processes.

Metabolic changes in response to drought conditions play a key role for drought adaptation in plants (Silvente et al. 2012). In the drought response network (Figure 4), we found some SNPs were associated with both drought-related gene expression phenotypes and metabolite level phenotypes. The genes containing the SNPs and the expressed genes provide candidates to analyze the genetic basis of metabolic changes in response to drought. Drought stress increases stearic acid (Júnior et al. 2008). SNP#56 resides in a gene encoding a cytochrome P450. It is associated with stearic acid concentration and *NAC1* (a drought-responsive TF) gene expression. Melezitose is found in the manna of many pine trees. During droughts, bees that collect manna from these trees produce honey containing elevated concentrations of melezitose (Purich 2017). SNPs #54 and #70 are associated with melezitose concentration and *RAP2.1* (a drought-responsive TF) gene expression. SNPs #54 and #70 reside in the genes encoding a cytochrome P450 and a MYB domain protein, respectively. It is probable that biosynthesis of melezitose in response to drought is under regulation of drought responsive genes.

This study is an attempt to compose networks for exploring the genetic basis of gene expression and metabolite level involved in complex biological processes. A total of 2.8 million SNPs were used to do association mapping, yet the numbers of investigated genes and metabolites are too limited to cover all the genes related to the

biosynthetic pathways. Numbers of genes related to ROS, PCD, terpenoid biosynthesis and phenylpropanoid pathway are too few to compose networks. In addition, gene expression and metabolite level were measured in different populations. If these data were to be measured with the same samples collected at the same time, the correlations between gene expression and metabolite level could be used to enrich the current networks. In the future, we wish to take advantage of the active development of transcriptome and metabolome profile technologies to improve the quantification of gene transcripts and metabolites.

Conclusion

We have identified a total of 1841 SNPs associated with 191 gene expression phenotypes and 524 SNPs associated with 53 metabolite level phenotypes. The identified SNPs reside in genes with a wide variety of functions. We constructed wood development and drought response networks and discovered key loci and genes that contribute to the biological processes. This work provides candidate genes to study the genetic basis of gene expression and metabolism involved in complex biological processes, and highlights the efficiency of using association-mapping-based networks to discover candidate genes involved in complex biological processes.

Supplementary Data

Supplementary Data for this article are available at *Tree Physiology* Online.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions

ML performed the sample collection and measurement, data analysis, and wrote the manuscript. KVK and CAL conceived and designed the study, coordinated the research and participated in the drafting of the manuscript. CMS helped with expression data analysis, interpretation and manuscript editing. All authors read and approved the final manuscript.

Funding

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.

Acknowledgements

We thank the Allele Discovery for Economic Traits in Pines 2 (ADEPT 2) project (National Science Foundation Grant DBI-0501763) for developing the population and providing the metabolite phenotyping data. We would like to thank Dr. Jill Wegrzyn

549 and the PineRefSeq Project (USDA National Institute of Food and Agriculture, Award
550 #2011-67009-30030) for providing the draft loblolly pine reference sequences and
551 exon annotation. We appreciate the Texas A&M Institute for Genome Sciences and
552 Society (TIGSS) for providing computational resources and system administration
553 support for the TIGSS HPC Cluster.

554

555 ## References

556

557 Ambawat S, Sharma P, Yadav NR, Yadav RC (2013) MYB transcription factor genes
558 as regulators for plant responses: an overview. Physiol Mol Biol Plants
559 19:307-321.
560 Belanger AE, Besra GS, Ford ME, Mikusová K, Belisle JT, Brennan PJ, Inamine JM
561 (1996) The embAB genes of Mycobacterium avium encode an arabinosyl
562 transferase involved in cell wall arabinan biosynthesis that is the target
563 for the antimycobacterial drug ethambutol. Proc Natl Acad Sci U S A
564 93:11919-11924.
565 Bossu J, Beauchêne J, Estevez Y, Duplais C, Clair B (2016) New Insights on Wood
566 Dimensional Stability Influenced by Secondary Metabolites: The Case of a
567 Fast-Growing Tropical Species Bagassa guianensis Aubl. PLoS one
568 11:e0150777.
569 Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007)
570 TASSEL: software for association mapping of complex traits in diverse
571 samples. Bioinformatics 23:2633-5.
572 Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB (2004) Nucleotide diversity
573 and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U S A
574 101:15255-60.
575 Burkhardt R, Kirsten H, Beutner F, Holdt LM, Gross A, Teren A, Tönjes A, Becker S,
576 Krohn K, Kovacs P (2015) Integration of genome-wide SNP Data and
577 gene-expression profiles reveals six novel loci and regulatory mechanisms
578 for amino acids and acylcarnitines in whole blood. PLoS Genet
579 11:e1005510.
580 Csardi G, Nepusz T (2006) The igraph software package for complex network
581 research. InterJournal, Complex Systems:1695. http://igraph.org/.
582 Cumbie WP, Eckert A, Wegrzyn J, Whetten R, Neale D, Goldfarb B (2011)
583 Association genetics of carbon isotope discrimination, height and foliar
584 nitrogen in a natural population of *Pinus taeda* L. Heredity 107:105-14.

585 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE,
586 Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project
587 Analysis G (2011) The variant call format and VCFtools. Bioinformatics
588 27:2156-8.
589 Denness L, McKenna JF, Segonzac C, Wormit A, Madhou P, Bennett M, Mansfield J,
590 Zipfel C, Hamann T (2011) Cell wall damage-induced lignin biosynthesis is
591 regulated by a reactive oxygen species-and jasmonic acid-dependent
592 process in Arabidopsis. Plant Physiol 156:1364-1374.
593 Diedrich G, Spahn CM, Stelzl U, Schäfer MA, Wooten T, Bochkariov DE, Cooperman
594 BS, Traut RR, Nierhaus KH (2000) Ribosomal protein L2 is involved in the
595 association of the ribosomal subunits, tRNA binding to A and P sites and
596 peptidyl transfer. EMBO J 19:5241-5250.
597 Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González -
598 Martínez SC, Neale DB (2010a) Patterns of population structure and
599 environmental associations to aridity across the range of loblolly pine
600 (*Pinus taeda* L., Pinaceae). Genetics 185:969-82.
601 Eckert AJ, Wegrzyn JL, Cumbie WP, Goldfarb B, Huber DA, Tolstikov V, Fiehn O,
602 Neale DB (2012) Association genetics of the loblolly pine (*Pinus taeda*,
603 Pinaceae) metabolome. New Phytol 193:890-902.
604 Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J,
605 DeFelice M, Lochner A, Faggart M (2002) The structure of haplotype
606 blocks in the human genome. Science 296:2225-2229.
607 González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007)
608 Association genetics in *Pinus taeda* L. I. wood property traits. Genetics
609 175:399-409.
610 Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population
611 structure with the assistance of sample group information. Mol Ecol
612 Resour 9:1322-32.
613 Júnior RRM, Oliveira MSC, Baccache MA, de Paula FM (2008) Effects of water
614 deficit and rehydration on the polar lipid and membranes resistance
615 leaves of Phaseolus vulgaris L. cv. Pérola. Braz arch biol technol
616 51:361-367.
617 Li Y, Pearl SA, Jackson SA (2015) Gene networks in plant biology: approaches in
618 reconstruction and analysis. Trends Plant Sci 20:664-675.
619 Lu M, Krutovsky KV, Nelson CD, Koralewski TE, Byram TD, Loopstra CA (2016)
620 Exome genotyping, linkage disequilibrium and population structure in
621 loblolly pine (*Pinus taeda* L.). BMC Genomics 17:730.
622 Lu M, Krutovsky KV, Nelson CD, West JB, Reilly NA, Loopstra CA (2017)
623 Association genetics of growth and adaptive traits in loblolly pine (Pinus
624 taeda L.) using whole-exome-discovered polymorphisms. Tree Genet
625 Genomes 13:57.
626 Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD (2016)

627 PANTHER version 11: expanded annotation data from Gene Ontology and
628 Reactome pathways, and data analysis tool enhancements. Nucleic Acids
629 Res:gkw1138.
630 Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene
631 function analysis with the PANTHER classification system. Nat Protoc
632 8:1551-1566.
633 Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers.
634 Trends Plant Sci 9:325-30.
635 Obata T, Witt S, Lisec J, Palacios-Rojas N, Florez-Sarasa I, Araus JL, Cairns JE,
636 Yousfi S, Fernie AR (2015) Metabolite profiles of maize leaves in drought,
637 heat and combined stress field trials reveal the relationship between
638 metabolism and grain yield. Plant Physiol:pp. 01164.2015.
639 Palle SR, Seeve CM, Eckert AJ, Cumbie WP, Goldfarb B, Loopstra CA (2011)
640 Natural variation in expression of genes involved in xylem development in
641 loblolly pine (Pinus taeda L.). Tree Genet Genomes 7:193-206.
642 Palle SR, Seeve CM, Eckert AJ, Wegrzyn JL, Neale DB, Loopstra CA (2013)
643 Association of loblolly pine xylem development gene expression with
644 single-nucleotide polymorphisms. Tree Physiol 33:763-74.
645 Pavy N, Namroud M, Gagnon F, Isabel N, Bousquet J (2012) The heterogeneous
646 levels of linkage disequilibrium in white spruce genes and comparative
647 analysis with other conifers. Heredity 108:273-284.
648 Phillips PC (2008) Epistasis--the essential role of gene interactions in the
649 structure and evolution of genetic systems. Nat Rev Genet 9:855.
650 Piscitelli A, Pezzella C, Giardina P, Faraco V, Sannia G (2010) Heterologous laccase
651 production and its role in industrial applications. Bioeng Bugs 1:254-264.
652 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure
653 using multilocus genotype data. Genetics 155:945-959.
654 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J,
655 Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for
656 whole-genome association and population-based linkage analyses. Am J
657 Hum Genet 81:559-75.
658 Purich D (ed) (2017) The inhibitor index: a desk reference on enzyme inhibitors,
659 receptor antagonists, drugs, toxins, poisons, biologics, and therapeutic
660 leads. CRC Press, Florida.
661 R Core Team (2017) R: A language and environment for statistical computing. R
662 Foundation for Statistical Computing, Vienna, Austria. URL
663 https://wwwr-projectorg/
664 Ro D-K, Arimura G-I, Lau SY, Piers E, Bohlmann J (2005) Loblolly pine
665 abietadienol/abietadienal oxidase *PtAO* (CYP720B1) is a multifunctional,
666 multisubstrate cytochrome P450 monooxygenase. Proc Natl Acad Sci U S
667 A 102:8060-8065.
668 Rosenberg E, Filer D, Zafriti D, Kindler S (1973) Aspartokinase activity and the

669 developmental cycle of Myxococcus xanthus. J Bacteriol 115(1):29-34.
670 Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B,
671 Wang S, Suver C (2008) Mapping the genetic architecture of gene
672 expression in human liver. PLoS Biol 6:e107.
673 Seeve CM. 2010. Gene expression and association analyses of stress responses in
674 loblolly pine (*Pinus taeda* L.). Texas A&M University.
675 Shin J-H, Blay S, McNeney B, Graham J (2006) LDheatmap: an R function for
676 graphical display of pairwise linkage disequilibria between single
677 nucleotide polymorphisms. J Stat Soft 16:Code Snippet 3.
678 Silvente S, Sobolev AP, Lara M (2012) Metabolite adjustments in drought tolerant
679 and sensitive soybean genotypes in response to water stress. PLoS One
680 7:e38554.
681 Slatkin M (2008) Linkage disequilibrium--understanding the evolutionary past
682 and mapping the medical future. Nat Rev Genet 9:477.
683 Wang X, Dalkic E, Wu M, Chan C (2008) Gene module level analysis: identification
684 to networks and dynamics. Curr Opin Biotechnol 19:482-491.
685 Wegrzyn JL, Liechty JD, Stevens KA, Wu LS, Loopstra CA, Vasquez-Gross HA,
686 Dougherty WM, Lin BY, Zieve JJ, Martinez-Garcia PJ, Holt C, Yandell M,
687 Zimin AV, Yorke JA, Crepeau MW, Puiu D, Salzberg SL, Dejong PJ, Mockaitis
688 K, Main D, Langley CH, Neale DB (2014) Unique features of the loblolly
689 pine (*Pinus taeda* L.) megagenome revealed through sequence annotation.
690 Genetics 196:891-909.
691 Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ (2007)
692 Linking metabolic QTLs with network and cis-eQTLs controlling
693 biosynthetic pathways. PLoS Genet 3:e162.
694 Westbrook JW, Resende MF, Jr., Munoz P, Walker AR, Wegrzyn JL, Nelson CD,
695 Neale DB, Kirst M, Huber DA, Gezan SA, Peter GF, Davis JM (2013)
696 Association genetics of oleoresin flow in loblolly pine: discovering genes
697 and predicting phenotype for improved resistance to bark beetles and
698 bioenergy potential. New Phytol 199:89-100.
699 Woodger FJ, Millar A, Murray F, Jacobsen JV, Gubler F (2003) The role of GAMYB
700 transcription factors in GA-regulated gene expression. J Plant Growth
701 Regul 22(2):176-184.
702 Wu Y, Deng Z, Lai J, Zhang Y, Yang C, Yin B, Zhao Q, Zhang L, Li Y, Yang C (2009)
703 Dual function of *Arabidopsis* ATAF1 in abiotic and biotic stress responses.
704 Cell Res 19(11):1279-1290.
705 Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical
706 power of nested association mapping in maize. Genetics 178:539-551.
707 Zhang Y, Brown G, Whetten R, Loopstra CA, Neale D, Kieliszewski MJ, Sederoff RR
708 (2003) An arabinogalactan protein associated with secondary cell wall
709 formation in differentiating xylem of loblolly pine. Plant Mol Biol
710 52:91-102.

711 ## Figure legends

712

713 **Figure 1**. Categorization of the transcription factor genes containing SNPs associated

714 with gene expression and metabolite level phenotypes. The gene expression

715 phenotypes were classified into different functional groups: wood-related,

716 disease-related, drought-related, reactive oxygen species (ROS)-related, terpenoid

717 biosynthesis, programmed cell death (PCD), and phenylpropanoid pathway. The

718 numbers above each bar represent the numbers of the identified SNPs associated with

719 gene expression or metabolite level phenotypes.

720

721 **Figure 2**. Pairwise linkage disequilibrium (LD) values for SNPs in the scaffold

722 tscaffold2867. The bottom vertex of each red triangle highlights the high LD values

723 for SNPs tscaffold2867_628232, tscaffold2867_651263 and tscaffold2867_755157

724 ($R^2 > 0.89$) located in the scaffold tscaffold2867.

725

726 **Figure 3**. Network comprised of the SNPs and their associated gene expression

727 (wood-related genes) and metabolite level phenotypes. Blue nodes represent SNPs.

728 Details of the SNPs and the genes containing them are presented in Table 2. The blue

729 node with a larger size represents the SNP that resides in a transcription factor (TF)

730 gene. Yellow nodes represent gene expression phenotypes. Pink nodes represent

731 metabolite level phenotypes. Grey edges represent associations between SNPs and

732 gene expression phenotypes. Red edges represent associations between SNPs and

733 metabolite level phenotypes. Purple edges represent SNP-SNP interactions that
734 significantly impact the phenotypes. Expressed genes in the network include:
735 arabinogalactan-protein and cell wall protein genes: *AGP1-6*; cell expansion genes:
736 *COB, KORRI*; cell wall related (resistance related) genes: *CslA1*; cellulose and callose
737 synthase genes: *CesA3*, *CslA2*, *CS-1343*; lignin biosynthesis enzyme genes: *4CL1*,
738 *C3H*, *CAD1*, *CCoAMT*, *COMT*, *Lac1-8*, *PAL1*, *TC4H*; α-tubulin gene: *αtub2*; wood
739 development enzyme genes: *BKACPS*, *BQR*, *Cellulase*, *EndChi*, *Importin*, *LP6*,
740 *PCBER*, *PLR*, *prxC2*, *SAH7*, *SPL*, *XET1*; wood development protein genes: *1CAB*-3A,
741 *NH-10*, *NH-9*, *RP-L2*; wood development TF genes: *SND1*, *AIP*, *APL*, *eIF-4A*, *FRA2*,
742 *KNAT4*, *KNAT7*, *LZP*, *MYB1*, *MYB4*, *MYB85*.
743
744 **Figure 4**. Network comprised of the SNPs and their associated gene expression
745 (drought-related genes) and metabolite level phenotypes. Blue nodes represent SNPs.
746 Details of the SNPs and the genes containing them are presented in Table 3. The blue
747 nodes with a larger size represent the SNPs that reside in transcription factor (TF)
748 genes. Pink nodes represent metabolite level phenotypes. Grey edges represent
749 associations between SNPs and gene expression phenotypes. Red edges represent
750 associations between SNPs and metabolite level phenotypes. Purple edges represent
751 SNP-SNP interactions that significantly impact the phenotypes. Expressed genes in
752 the network include: drought signaling genes: *ABI1*, *NCED*, *RPK1*;
753 drought-responsive TF genes: *NAC1*, *RAP2.1*, *RAP2.4*, *ATAF-1*; late embryogenesis

754 abundant protein genes: *PtEMB3*-4; phenylpropanoid pathway gene: *ANR*.