New Phytologist

Research

Genomic architecture of complex traits in loblolly pine

Amanda R. De La Torre[1,2], Daniela Puiu[3], Marc W. Crepeau[4], Kristian Stevens[4], Steven L. Salzberg[3,5], Charles H. Langley[4] and David B. Neale[2]

[1]School of Forestry, Northern Arizona University, 200 E. Pine Knoll Drive, Flagstaff, AZ 86011, USA; [2]Department of Plant Sciences, University of California-Davis, One Shields Avenue, Davis, CA 95616, USA; [3]Center for Computational Biology, Johns Hopkins University, 1900 E. Monument St, Baltimore, MD 21205, USA; [4]Department of Evolution and Ecology, University of California-Davis, One Shields Avenue, Davis, CA 95616, USA; [5]Department of Biomedical Engineering, Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD 21205, USA

Author for correspondence:
Amanda R. De La Torre
Tel: +1 530 574 3385
Email: Amanda.de-la-torre@nau.edu

Received: *25 July 2018*
Accepted: *6 October 2018*

New Phytologist (2019) **221:** 1789–1801
doi: 10.1111/nph.15535

Key words: balancing selection, complex traits, genome-wide association studies (GWAS), loblolly pine, metabolites, mutation-selection balance, pitch canker, xylem development genes.

Summary

- Dissecting the genetic and genomic architecture of complex traits is essential to understand the forces maintaining the variation in phenotypic traits of ecological and economical importance.
- Whole-genome resequencing data were used to generate high-resolution polymorphic single nucleotide polymorphism (SNP) markers and genotype individuals from common gardens across the loblolly pine (*Pinus taeda*) natural range. Genome-wide associations were tested with a large phenotypic dataset comprising 409 variables including morphological traits (height, diameter, carbon isotope discrimination, pitch canker resistance), and molecular traits such as metabolites and expression of xylem development genes.
- Our study identified 2335 new SNP × trait associations for the species, with many SNPs located in physical clusters in the genome of the species; and the genomic location of hotspots for metabolic × genotype associations.
- We found a highly polygenic basis of quantitative inheritance, with significant differences in number, effects size, genomic location and frequency of alleles contributing to variation in phenotypes in the different traits. While mutation-selection balance might be shaping the genetic variation in metabolic traits, balancing selection is more likely to shape the variation in expression of xylem development genes. Our work contributes to the study of complex traits in nonmodel plant species by identifying associations at a whole-genome level.

Introduction

Significant progress has been made towards understanding the evolutionary forces that shape neutral genetic variation among and within natural populations of plant species. Although of widespread interest, the understanding of how genomic variation is maintained for complex traits is still very limited (Mitchell-Olds *et al.*, 2007; Josephs *et al.*, 2017). If the underlying polymorphisms are largely deleterious, a balance between their creation by mutation and their elimination by natural selection, will maintain the variation in complex traits (Lande, 1975; Barton & Keightley, 2002). Alternatively, natural selection will maintain the variation in individuals by balancing selection or throughout the entire species by local adaptation (Turelli & Barton, 2004; Charlesworth, 2006). Larger phenotypic and genotypic datasets combined with powerful analysis such as GWAS, QTL mapping and high-density linkage mapping may allow us to evaluate the forces shaping the genetic and genomic architecture of complex traits.

In plants, GWAS studies have largely been used in model or crop species and its use in natural populations of nonmodel species is relatively new. Due to the strong effects of artificial selection and domestication in most crop species studied to date, the amount and structure of the genomic variation underlying complex trait variation can be qualitatively different from that in outcrossing undomesticated plant species (such as loblolly pine). Conifer tree species differ from many studied plant species by other important features that are likely to shape variation in complex traits, including their mating system (outcrossing vs selfing), ploidy (diploid vs polyploid), population sizes (large vs small), generation-time (long-lived vs short-lived) and mutation rates (low vs high; De La Torre *et al.*, 2014, 2017).

The dissection of complex traits in conifer trees began more than 20 years ago with QTL mapping of traits of economic importance (Groover *et al.*, 1994). With the advent of the genomic era, association mapping became a promising tool due to the undomesticated status of conifer trees, outcrossing mating system and rapid decay of linkage disequilibrium (Neale & Kremer, 2011). This made it possible to detect larger numbers of polymorphisms in close proximity to QTL or even QTL themselves, although effect sizes are reported to be smaller than those

detected by QTL mapping (Neale & Savolainen, 2004; Neale & Kremer, 2011; Hall *et al.*, 2016; Plomion *et al.*, 2016). With the absence of reference genomes, these studies focused only on candidate genes or specific regions (although see Fuentes-Utrilla *et al.*, 2017; Lu *et al.*, 2017). As a result, most of the associations explained a very small proportion of the genetic variation in complex traits (Plomion *et al.*, 2016). The recent sequencing of reference genomes of a few commercially important conifers brings an opportunity to fill this gap. However, the dissection of the genetic architecture of complex traits is still challenging due to the fragmented nature of the reference genomes, incomplete gene annotation, and the absence of physical and high-density linkage maps (De La Torre *et al.*, 2014; Neale *et al.*, 2017). Also, due to the highly polygenic basis of complex traits and rapid decay of linkage disequilibrium, large mapping population sizes are needed to capture a large proportion of the genetic variation, especially when using GWAS instead of QTL mapping (Hall *et al.*, 2016).

Due to its high economic importance in the southern United States, loblolly pine (*Pinus taeda*) has been the focus of several studies aimed at describing the phenotypic variation of complex traits and their association with genotypes (Gonzalez-Martinez *et al.*, 2007, 2008; Eckert *et al.*, 2010a,b, 2012; Quesada *et al.*, 2010; Cumbie *et al.*, 2011; Palle *et al.*, 2011, 2013; Lu *et al.*, 2017). All of these studies have suggested a polygenic basis for complex traits, with a large number of genes of small effect sizes, additive effects, and the contributions of both coding and noncoding portions of genes (Eckert *et al.*, 2012). Although these results are largely consistent with the theoretical predictions for complex traits, knowledge of the evolutionary forces that maintain genome-wide variation for complex traits is nonexistent for conifer trees.

In this study, we use genome-wide single nucleotide polymorphisms (SNPs) derived from whole-genome resequencing of widely distributed individuals of the species to run a GWAS in combination with a newly generated 26k SNP high-density linkage map (A. R. De La Torre & D. B. Neale, unpublished). With these, we aim to dissect the genetic and genomic architecture of a large number of 409 morphological and molecular traits, and to understand the evolutionary forces that shape the genetic variation for complex traits in the species. Our work contributes to the study of adaptive traits in forest trees by identifying significant genome-wide associations in a nonmodel plant species.

Materials and Methods

Sample collection and DNA isolation for whole-genome resequencing

Seeds from 10 loblolly pine individuals spanning all the species' natural distribution were collected for genomic analysis (Fig. 1a). Before DNA extraction, seeds were soaked in water at room temperature for 4 d, then haploid megagametophytes were dissected from each seed. DNA was extracted from megagametophytes with the Qiagen DNeasy Mini-prep Plant kit and quantified using picogreen on a Qubit fluorometer.

Whole-genome resequencing

Libraries were constructed using the Illumina's TruSeq Nano DNA Library Prep Kit, according to the Illumina's Sample Preparation Guide (Illumina Part #15041110 Rev. B). Preamplification steps included DNA fragmentation (200 ng starting material and a 550-bp target insert size), followed by end repair and size selection of fragments, adenylation of 3′ ends, and ligation of adapters. Eight cycles of polymerase chain reaction (PCR) enrichment were performed. Barcoded libraries were combined into normalized pools and sequenced to > 10× coverage on an Illumina HiSeq

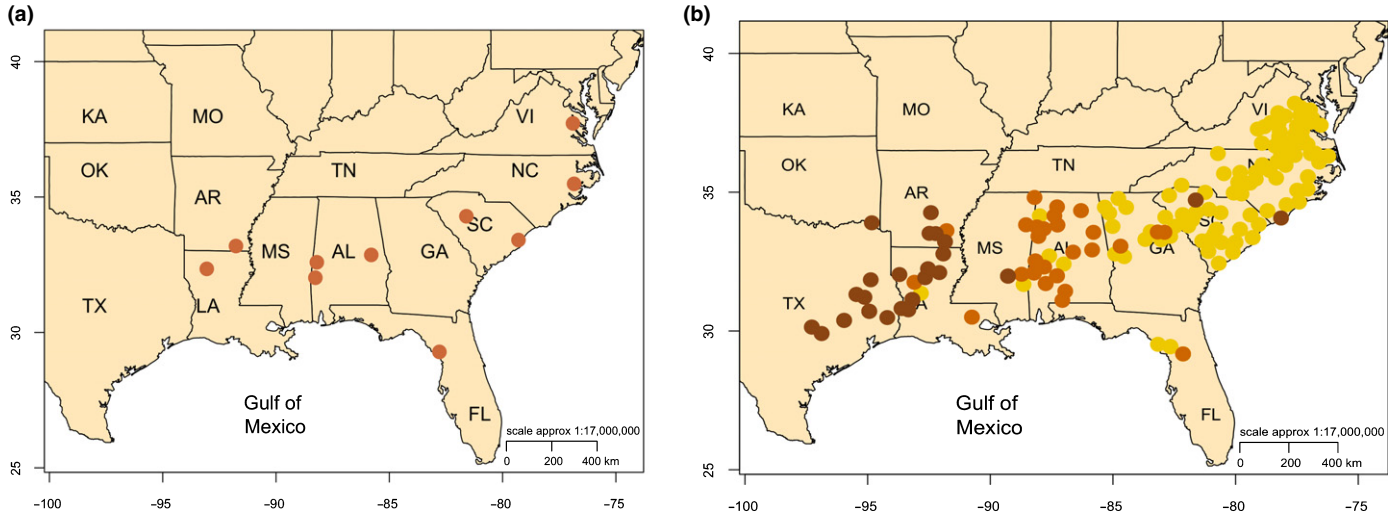

Fig. 1 Geographical distribution of loblolly pine individuals and populations sampled in this study. (a) Map of whole-genome resequenced individuals; (b) map of individuals clustered in three populations based on the results of the principal component analysis (PCA) with 87k single nucleotide polymorphism (SNP) markers.

3000 (Illumina Inc., San Diego, CA, USA) instrument using 150-bp paired-end reads. Sequencing took place at the Genome Center of the University of California, Davis, California.

SNP calling

Raw reads from whole-genome resequencing data from 10 loblolly pine individuals were aligned to the loblolly pine reference genome Lp.v.2.0 (GenBank accession GCA_000404065.3; Zimin *et al.*, 2017) using Bowtie2 v.2.2.9 (Langmead & Salzberg, 2012). SNP calling was done using Samtools v.1.3.1, followed by Bedtools v.2.25.0 and Bfctools v.1.3.1 (Li *et al.*, 2009; Li, 2011). Default parameters were used in Samtools and Bedtools. Filtering criteria included the removal of SNPs with a quality < 20, depth of coverage < 8, mapping quality = 0 or less, and indels. More details on the SNP calling and SNP filtering criteria and selection can be found in Supporting Information Methods S1.

Sample collection, DNA extraction and SNP genotyping

Needles from 377 outcrossing, unrelated individuals from the ADEPT2 common garden located in Mississippi, southeast United States were collected (Fig. 1b; Eckert *et al.*, 2012). DNA was extracted from needle tissue with a lab protocol that included 1 d of tissue lysis and incubation at 96°C, followed by several steps of precipitation and filtering using the Qiagen DNeasy Mini-prep Plant kit with an Eppendorf automated pipetting workstation. DNA was quantified using picogreen on a Qubit Fluorometer. From the 5.2 M SNPs obtained from the SNP calling of the resequencing data, 635k SNP markers were selected for genotyping. Affymetrix used *in silico* design scores to maximize the number of markers for genotyping that will provide high conversion rates. Samples were genotyped using an Affymetrix Axiom myDesign species-specific and customized SNP array comprising 635k SNP markers. The Axiom Analysis Suite v.3.1 (Thermo Fisher Scientific Inc., Waltham, MA, USA, 2017) was used to obtained genotyping statistics for all samples and SNPs. Samples with a dish QC (dQC) value greater than or equal to 0.82, and QC call rate of 97% obtained from the analyses, were kept for further analyses. SNPs were kept when the cluster call rate (CR) was equal or higher than 97%.

Population structure

Population structure in the dataset was evaluated with a principal component analysis (PCA) in Adegenet v.2.0.1 R package (Jombart, 2008; Jombart & Ahmed, 2011). Input files for Adegenet were obtained using the -recode function in Plink 1.9 (Chang *et al.*, 2015). In addition, the Python2.x FastStructure algorithm based on a variational framework was used for posterior inference of K clusters (Raj *et al.*, 2014) using input files in *bed*, *bim*, and *fam* format obtained using the -make-bed function in Plink v.1.07 (Purcell *et al.*, 2007). Models in FastStructure were replicated 10 times with K from 1 to 10 using the default prior; seeds for random number generators were changed for each run. The chooseK.py python script in FastStructure was used to estimate the model complexity that maximizes marginal likelihood and the model components were used to explain structure in the data.

Genotype × phenotype genome-wide association study

We used a large phenotypic data set comprising 409 morphological and molecular phenotypes (Table S1). Morphological traits included height at age 3, diameter at breast height (dbh) at age 3, drought resistance (carbon isotope discrimination measured from foliage after the end of the second growing season), and disease resistance (pitch canker resistance measured as lesion length at 4, 8 and 12 wk after inoculation). Molecular traits included 292 metabolites (extracted from xylem tissue using gas chromatography and mass spectrometry), and gene expression of 111 xylem development genes. All phenotypic values with the exception of height and dbh were previously reported (Quesada *et al.*, 2010; Cumbie *et al.*, 2011; Palle *et al.*, 2011; Eckert *et al.*, 2012). Estimates of clonal phenotypic values for pitch canker resistance and carbon isotope discrimination were obtained using best linear unbiased predictions (BLUP) implemented in lme4 R package v.1.1-14. Clonal least-square means were also adjusted using mixed linear models for the metabolome data.

To identify significant genes explaining phenotypic variation, a GWAS analysis was carried on for each of the 409 phenotypes with 87 825 SNPs with compressed mixed linear model (Zhang *et al.*, 2010) implemented in the gapit R package (Lipka *et al.*, 2012). Principal components were used as co-variants to account for population structure. Locations of SNPs in the genome of the species were obtained from our newly constructed 26k high-density linkage map for loblolly pine (A. R. De La Torre & B. B. Neale, unpublished). Manhattan plots were built with the qqman R package (Turner, 2014). SNP functional annotations were obtained from the annotated genome of loblolly pine v.2.01 in TreeGenes (http://treegenesdb.org/FTP/Genomes/Pita/v2.01/annotation/), from aligning against the full NCBI Nucleotide collection database using Blastn 2.8.0, and from aligning against the nonredundant protein sequences database using Blastx 2.8.0 (e value < $1e^{-10}$; Zheng *et al.*, 2000). We tested the presence of false positives in our results, by assessing the enrichment of likely functional variants (nonsynonymous) versus neutral variants (synonymous) in all SNPs in genes and exons. Nucleotide sequences with alternative alleles were translated and then aligned to the corresponding proteins using Blastx (-num_alignments 1, -num_threads 24, -outfmt 6); differences in protein sequences were accounted as nonsynonymous SNP variants. We also estimated allelic effects (increase in phenotype when favorable allele is in homozygous state) for each SNP in each of the phenotypes. A positive allelic effect sign indicates that the favorable allele is the second in alphabetic order.

Phenotype × environment associations

In total, 282 monthly, seasonal, and annual variables were obtained from climate normal data from 1961 to 1990 from

ClimateNA v.5.41 (Wang *et al.*, 2016). Potential correlations among all variables in the Phenotypic and Environmental data sets were evaluated using the rcorr function in the R package Hmisc v.4.1-1 (http://biostat.mc.vanderbilt.edu/Hmisc). Geographical variables (latitude, longitude and elevation) were also analyzed. In total, 234 955 correlations were evaluated and corrected for multiple testing using the Bonferroni–Holmes correction ($P < 0.05$).

Data availability

All Illumina whole-genome resequencing data (10 libraries, 460 files) are available at the NCBI Sequence Read Archive (SRA) under bioproject PRJNA174450 (https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA174450&go=go).

Results

SNP calling and SNP genotyping

In total, 455M SNPs were identified that occurred in any of the whole-genome resequenced individuals (Fig. 1a). From these, 5.2 M SNPs containing a combination of transcript-aligned and not transcript-aligned hitting one or more resequenced individuals were selected. In total, 635 453 best-scoring SNPs were included in the SNP array for genotyping. The SNPs in this array were distributed in 132 093 scaffolds that represent *c.* 9% (1.94Gb) of the total genome size of loblolly pine v.2.0 (Neale *et al.*, 2014; Zimin *et al.*, 2017).

After SNP array genotyping and initial filtering, 569 542 (95.3%) SNPs and 359 (95.25%) samples passed the quality control criteria. In total, 324 915 SNPs were found to be in Hardy–Weinberg equilibrium ($P < 0.05$) when estimated using the Axiom Analysis Suite v.3.1 (Thermo Fisher Scientific Inc., 2017). From all SNPs, 6.7% were monomorphic with high-resolution, and 22.5% were polymorphic markers with cluster properties above the threshold. From the polymorphic markers, we kept 84 744 SNPs for further analysis. In addition, we added 3081 gene-based SNPs previously reported in Eckert *et al.*, 2010a,b, resulting in 87 825 SNPs.

Population structure

The results of the PCA analysis suggest the presence of three major genetic clusters in the dataset that extend longitudinally across the species' natural range. The eastern cluster is composed by populations in Virginia, North Carolina, South Carolina, Georgia and Florida; the center cluster contains populations in Alabama and Mississippi; and the western cluster contains populations in Texas, Arkansas and Louisiana (Fig. 1b). Posterior inference of clusters based on variational Bayesian framework implemented in FastStructure suggested $K = 2$ better explains the genetic structure of the species. When $K = 2$, the center and eastern populations are differentiated from the western populations, suggesting that the Mississippi river is the major barrier for gene flow. This population structure is coincident with previous smaller-scale studies in the species using SNPs and simple sequence repeats (SSRs) (Eckert *et al.*, 2010a,b) and also with more recent ones (Lu *et al.*, 2017), suggesting a dual Pleistocene refugia model might have shaped the genetic structure of loblolly pine.

Genotype × phenotype genome-wide association study

In total, 2335 significant associations (FDR-Adjusted *P*-value < 0.05) were found among 1726 SNPs and 111 phenotypes, including carbon isotope discrimination, pitch canker disease resistance at 8 and 12 wk, 97 metabolites and 12 expression of xylem development genes (Tables S2–S5). From these SNPs, 801 came from coding sequences, and either match transcripts, exons or genes in the loblolly pine reference genome Lpv.2.0, or in the loblolly pine reference transcriptome (August 2016 version). In total, 680 SNPs were physically located in linkage groups, and from these, 175 were located in our newly built 26k loblolly pine linkage map (A. R. De La Torre & D. B. Neale, unpublished data).

Coincident with previous GWAS and QTL mapping studies in conifer species, we found a largely polygenic genomic basis of inheritance of complex traits in loblolly pine. Individual traits were associated with a variable number of SNPs ranging from 1 to 448 SNPs per trait, suggesting that very few of the traits in this study have single-gene inheritance. Single SNPs associated with several traits were observed especially among metabolites, suggesting the pleiotropic nature of these SNPs. Most of the SNPs were found to be associated to two or three traits, however we found two SNPs associated with seven different metabolites (AX-173181025 from linkage group 7, and AX-173182344 from linkage group 8). For example, 221 individual SNPs were found to be associated with either two or three of the metabolites met_296133, met_216839, and met_338849. SNPs associated with these metabolic traits were found to be co-located or in close proximity in linkage groups 3, 4, 5 and 8. Among the SNPs associated with three metabolites were SNP 2-7725-01-553, a β-galactosidase 8 gene located in linkage group 6, AX-172848823 (nicalin 1-like, linkage group 3), AX-172878034 (linkage group 12), and AX-172896656 (linkage group 10). Among the individual SNPs associated with both met_296133 and met_216839 were SNP 2-5139-01-378 (calmodulin gene), SNP 2-6413-01-182 (WD-repeat protein), AX-172779046 (linkage group 4, PITA_00552 gene), AX-172796879 (linkage group 5), and AX-172802581 (linkage group 9). Similarly, 39 individual SNPs were associated with both metabolites threonine and phenylalanine, including SNP AX-166030910 (linkage group 1, unknown mRNA), AX-172783974 (PITA_00901 gene, Pt1m2 early response to drought erd3) and AX-172814594 (linkage group 9). Only one SNP was found to be associated to both metabolites and gene expression traits. This was the case for SNP AX-173395058 associated with metabolites met_296133 and met_216839, and the expression of the enzyme β-ketoacyl-ACP-synthetase I-2 (BKACPS; Table S4).

The distribution of associated SNPs across the genome of loblolly pine varied according to the type of trait studied. For

example, SNPs associated with metabolic traits seem to be widely distributed across the genome, showing their presence in each of 12 linkage groups (Fig. 2). Most metabolic traits showed several candidate mQTLs located in different linkage groups. Although of widespread distribution, clusters of mQTL (67 associations between 34 SNPs and 25 metabolic traits) were observed in 6 of the 12 linkage groups. Clusters were located in linkage group 2 position 189.47–193.4 cM; linkage group 3 position 100.08–103.44 cM; linkage group 4 position 83.65 cM; linkage group 5 position 48.76–49.24 cM; linkage group 8 position 153.15 cM; and linkage group 10 positions 133.22–136.03 cM (Fig. 3; Table S3). Functional annotation of SNPs suggests that SNPs clustered in linkage group 4 matched transcripts coding for E3 ubiquitin ligases. As we lacked functional annotation of several other clustered SNPs, we could not tell if there were any other functional connections among SNPs in any of the other clusters.

SNPs associated with any of the expression of xylem development traits were found to occur in the same linkage group and are thought to be clustered (within proximate location), although we were unable to estimate the precise location (cM) within linkage groups of several of these SNPs. For example, SNPs associated with the expression of UDP-glucose pyrophosphorylase (Pyro) co-occur in linkage group 7. Xyloglucan endotransglycosylase-3 associated SNPs co-occur in linkage group 8, calose synthase 3 in linkage group 11, cellulose synthase 2 in linkage group 6, and horseradish peroxidase C2 in linkage group 4 (Fig. 4). The exceptions are the SNPs associated with the expression of α-tubulin 1, which do not seem to be clustered but instead are distributed in four different linkage groups (4, 2, 9 and 8; Table S5).

In addition, the majority (8 out of the 13) of SNPs with known linkage group and associations with pitch canker disease resistance at 8 and 12 wk were located in the same scaffold

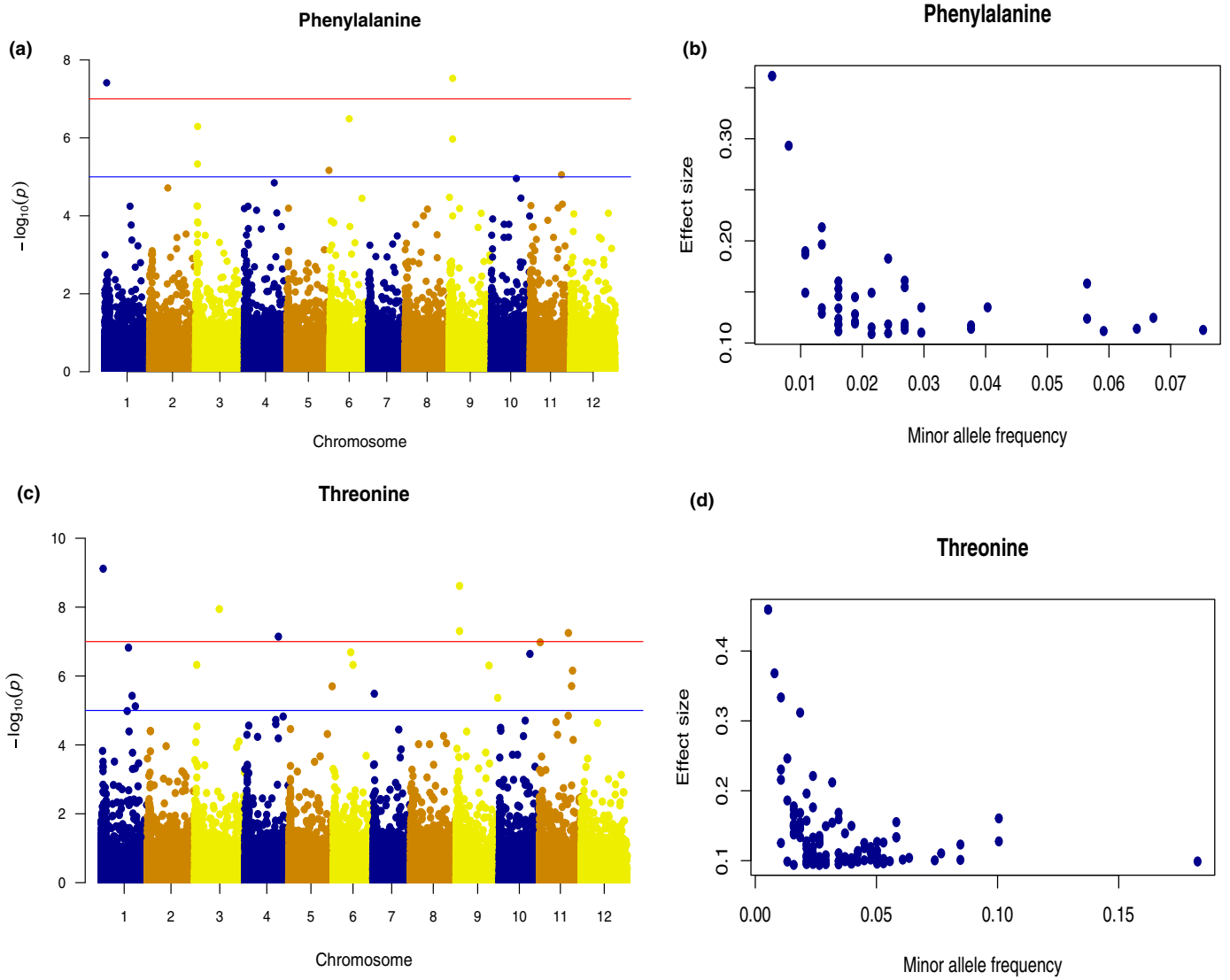

Fig. 2 Manhattan plots showing significant associations in metabolites in loblolly pine. (a) Phenylalanine and (c) threonine; and the relationship between minor allele frequency and effect size for each trait (b, d). Blue horizontal lines indicate a threshold of $\log_{10}(1e^{-05})$; red lines indicate a threshold of $\log_{10}(1e^{-07})$. The latter was used for significant marker-trait associations.

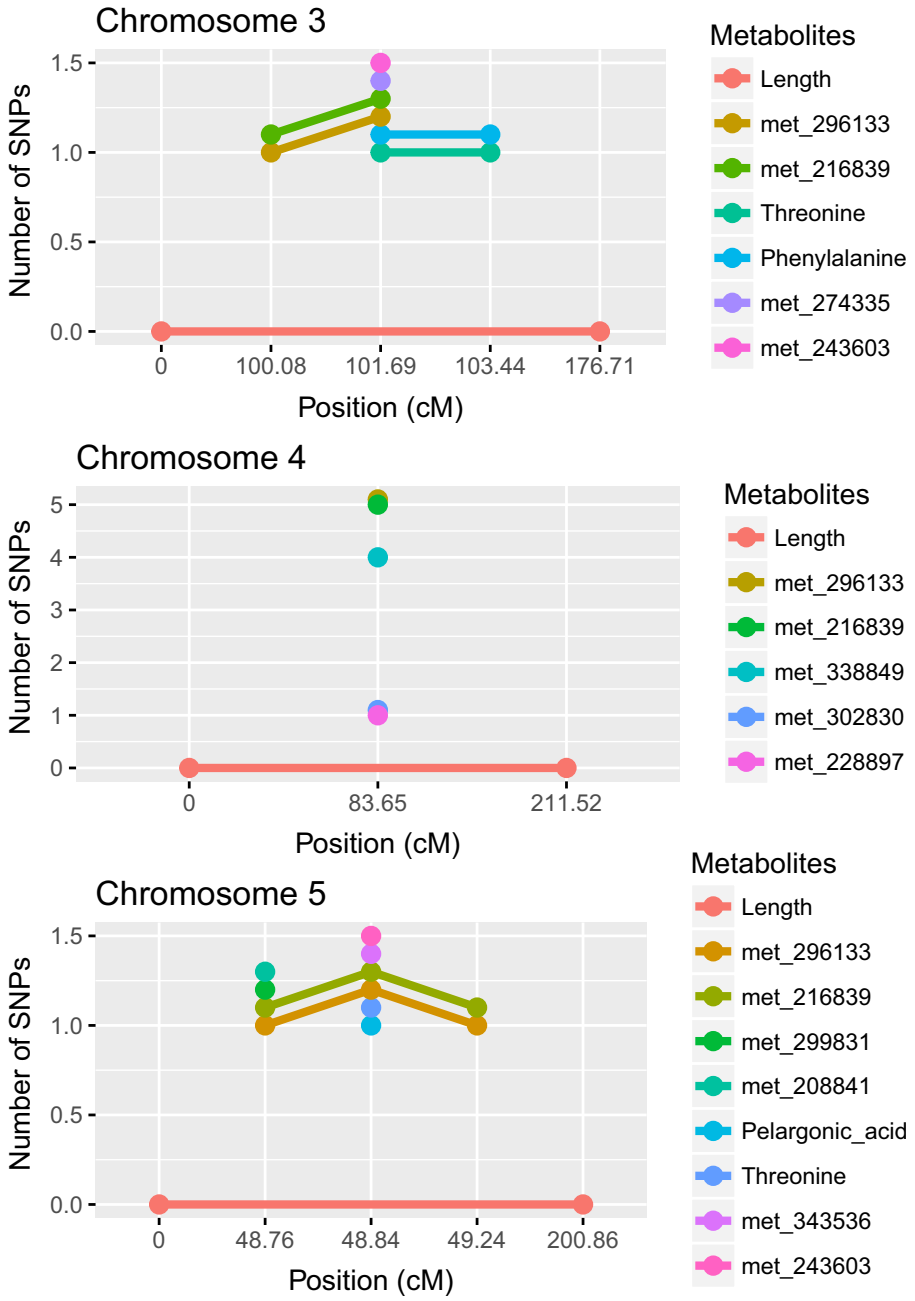

Fig. 3 Genomic locations of hotspots for metabolic QTLs (mQTLs) and number of single nucleotide polymorphisms (SNPs) associated with each metabolite in loblolly pine. Length indicates the total linkage group length.

(scaffold 109981) and linkage group 3 (Fig. 5; Table S2). Identified as variants in genes in the ABC transporter family, these SNPs had strong allelic effects in heterozygote genotypes, producing large increments in lesion lengths after inoculation of pitch canker disease (Fig. 5d). Recessive homozygotes were not found for these SNPs, suggesting either these susceptible individuals might not have been able to survive into adulthood or their presence is extremely rare in nature. Similar allelic effects were found in other significant SNPs with unknown linkage group position or missing annotations (Table S6). Allelic effects for metabolic and expression traits were much lower than for pitch canker disease resistance (Tables S7, S8).

Minor allele frequencies (maf) of significant SNPs varied from 0.005 to 0.49. When adjusting for minor allele frequency to higher than 0.03, the number of significant associations is reduced to 862. When comparing the minor allele frequencies between morphological and molecular traits, we found different patterns in the number of SNPs associated, effect sizes and their minor allele frequencies. Significant morphological traits, carbon isotope discrimination and pitch canker resistance at 8 and 12 wk were associated with a small group of SNPs (2 per carbon isotope and 18 per pitch canker), a median maf equal to 0.021, and small effect sizes from 0.09 to 0.15 (mean = 0.10, SD = 0.015). Significant

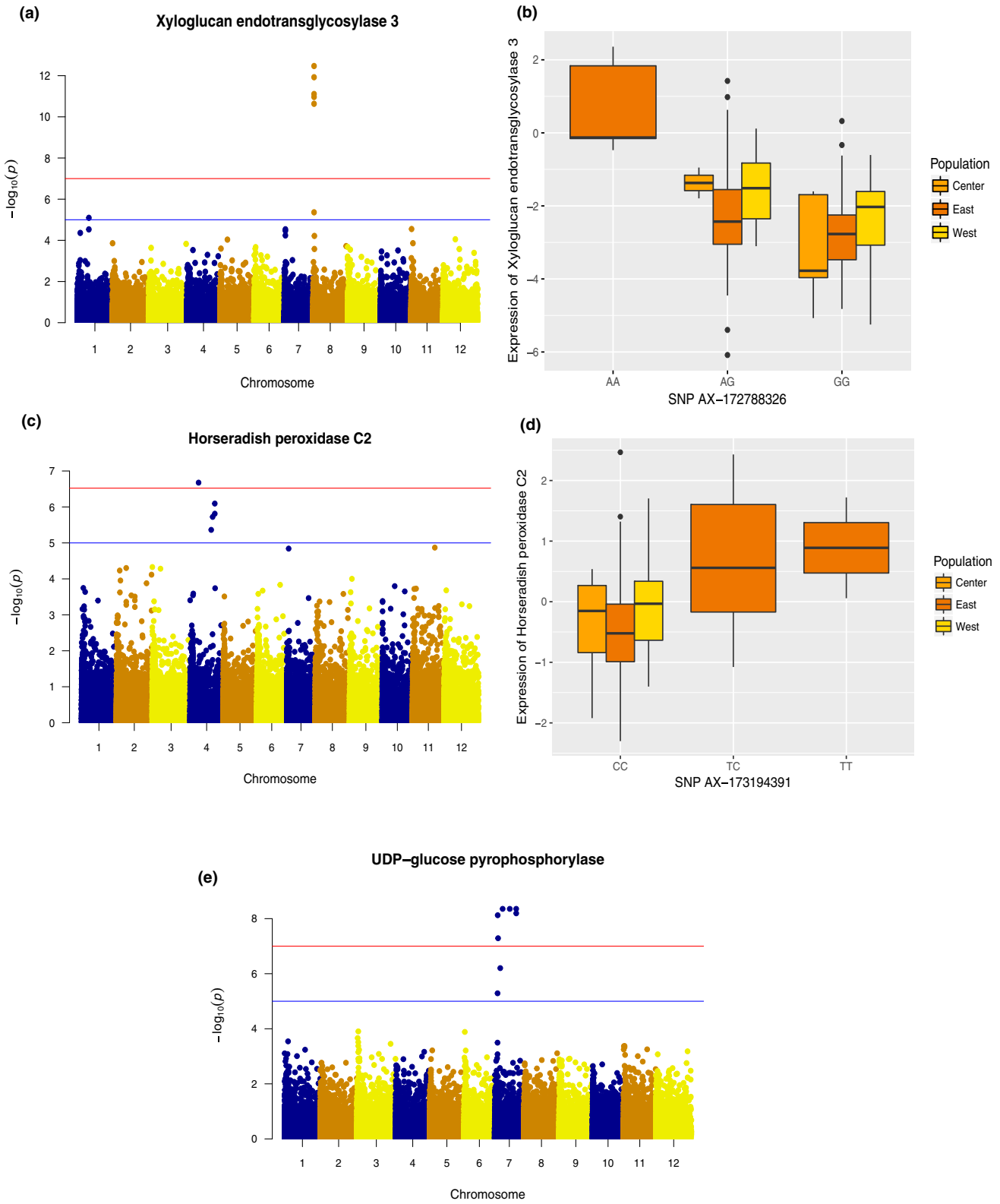

Fig. 4 Significant single nucleotide polymorphism (SNP) associations with expression of xylem development genes in loblolly pine. Manhattan plot (a) and expression variation (b) in xyloglucan endotransglycosylase 3; Manhattan plot (c) and expression variation (d) in horseradish peroxidase C2; and Manhattan plot for UDP-glucose-pyrophosphorylase (e). Blue horizontal lines in Manhattan plots indicate a threshold of $\log_{10}(1e^{-05})$; red lines indicate a threshold of $\log_{10}(1e^{-07})$. The latter was used for significant marker-trait associations. Boxplots in (b) and (d) indicate the variation in expression values, in which horizontal lines are the median values, and black circles represent outliers.

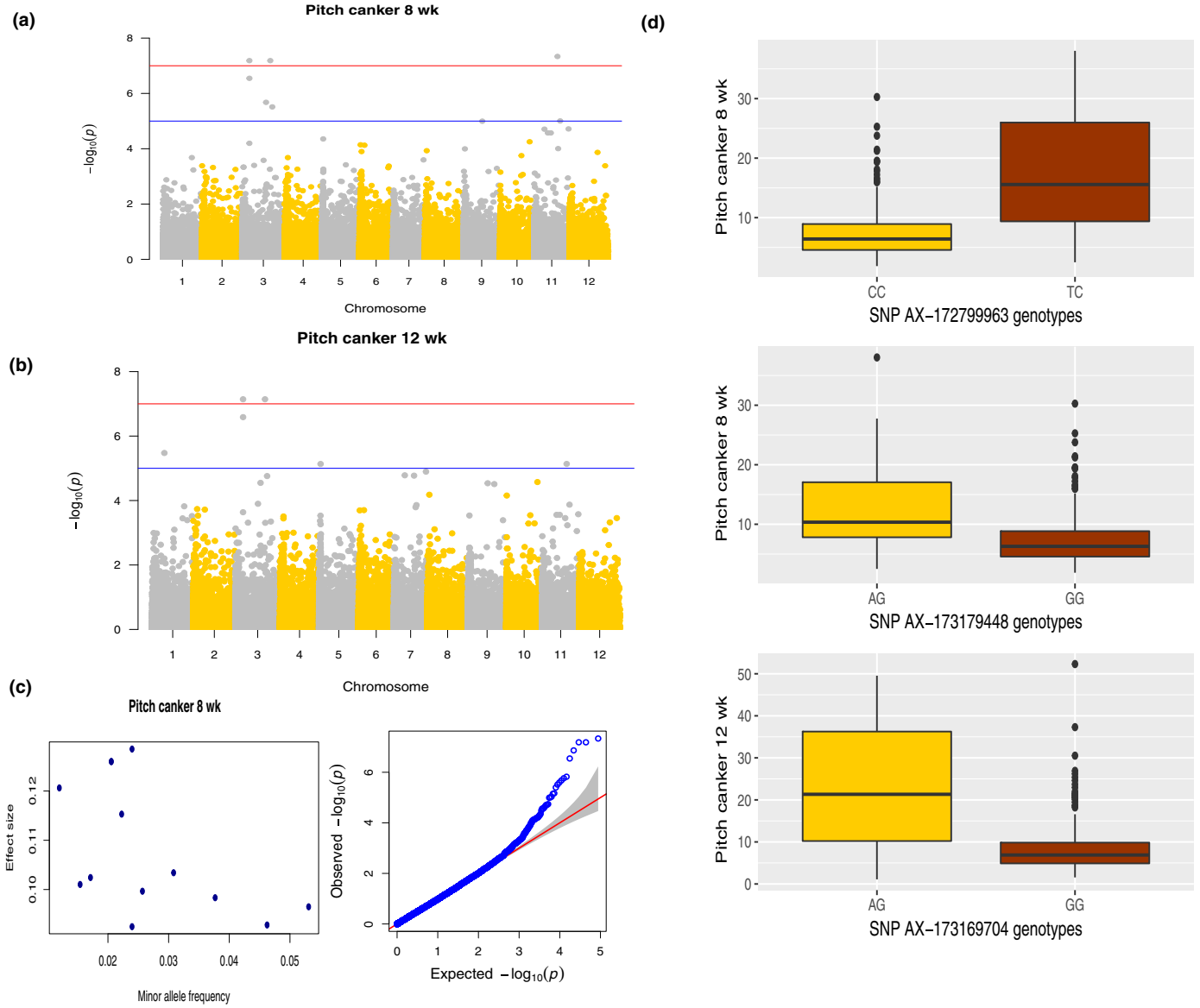

Fig. 5 Genome-wide association study (GWAS) results of pitch canker resistance in loblolly pine. Manhattan plots showing significant associations for lesion length after 8 wk (a) and lesion length after 12 wk (b); relationship between minor allele frequency and effect size, and QQ-plot for lesion length after 8 wk (c); significant allelic effects for three SNPs co-occurring at linkage group 3 (d). Blue horizontal lines in Manhattan plots indicate a threshold of $\log_{10}(1e^{-05})$; red lines indicate a threshold of $\log_{10}(1e^{-07})$. The latter was used for significant marker-trait associations. Boxplots indicate the variation in pitch canker lesion lengths, in which horizontal lines are the median values and black circles represent outliers. In the QQ-plot (c, right panel), negative logarithms of the *P*-values from the GWAS models were plotted against their expected value under the null hypothesis of no marker-trait association (vertical red line). Grey area shows the 95% confidence interval under the null hypothesis of no marker-trait association.

molecular traits such as metabolites were associated with a much variable and larger number of SNPs (1 to 448 SNPs per trait) with a median maf equal to 0.023, and small to large effect sizes from 0.09 to 0.56 (mean = 0.16, SD = 0.06). By contrast, molecular traits such as expression of xylem development genes were associated with also a small group of SNPs (1 to 17 SNPs per trait), a larger median maf equal to 0.2156, and small to moderate effect sizes from 0.09 to 0.389 (mean = 0.15, SD = 0.04).

We tested for a correlation between effect size (proportion of the variance explained) and SNP minor allele frequency in all 111 traits showing significant associations with at least 10 SNPs. Our results show that effect size is inversely correlated (*P*-value < 0.01) with maf in metabolites met_216839, met_217866, met_338849, met_299831, met_243603, met_318115, threonine, met_281189, glyceric acid, met_299833, phenylalanine, met_299711, met_238943, met_235011, isoleucine, and six other metabolites with *P*-values between 0.02 and 0.0469 (Fig. 2b,d; Table S9). No significant associations were found between effect size and maf in any other trait besides metabolites. However, it is clear that SNPs with intermediate maf have the higher effect sizes in SNPs correlated with expression of xylem development genes (Fig. S1; Table S5). In the case of SNPs associated with pitch canker, most of the GWAS significant alleles

had maf < 0.05, suggesting rare alleles play an important role (Fig. 5; Table S2).

From the significant SNPs located in genes and exons, we found that 73% were nonsynonymous. From the ones that were synonymous – potentially false positives – 68% had minor allele frequencies lower than 0.03. However, there was not significant association between synonymous/nonsynonymous variants and mean allele frequencies for all significant SNPs in genes and exons.

Phenotype × environment associations

After correcting for multiple testing using Bonferroni–Holmes correction ($P < 0.05$), we found 41 significant correlations among phenotypic traits and environmental traits, or among phenotypic traits. All significant associations came from expression of xylem development genes. For example, the expression of the enzyme myo-inositol-1-phosphate synthase was negatively correlated with precipitation in August ($R^2 = -0.322$, $P < 0.001$), and positively correlated with Hargreaves climatic moisture deficit (CMD08) in August ($R^2 = 0.33$, $P < 0.001$) and summer ($R^2 = 0.32$, $P < 0.001$); as well as positively correlated with 13 other expression traits (Table S10). Expression of secretory protein (SPL) was negatively correlated with radiation during the summer (Rad_sm; $R^2 = -0.32$, $P < 0.001$) and August (Rad08; $R^2 = -0.35$, $P < 0.001$); and positively correlated with the expression of the enzyme xyloglucan xylosyl transferase 5. Expression of arabinogalactan 5D was negatively correlated with precipitation in June ($R^2 = -0.32$, $P < 0.001$). Expression of enzyme xyloglucan endotransglycosylase 2 (XET-2) was negatively correlated with precipitation in August ($R^2 = -0.34$, $P < 0.001$) and summer ($R^2 = -0.329$, $P < 0.001$), and radiation in May ($R^2 = 0.33$, $P < 0.001$) and spring ($R^2 = 0.326$, $P < 0.001$); and positively correlated with Hargreaves climatic moisture deficit during the summer ($R^2 = 0.32$, $P < 0.001$) and nine other expression traits (Table S10). Finally, expression of MADS box protein AGL2 was positively correlated with Hargreaves climatic moisture deficit in May ($R^2 = 0.33$, $P < 0.001$), Radiation in March ($R^2 = 0.32$, $P < 0.001$), May ($R^2 = 0.33$, $P < 0.001$) and spring ($R^2 = 0.326$, $P < 0.001$), and expression of lignin biosynthesis enzyme laccase 6 ($R^2 = 0.315$, $P < 0.001$); and negatively correlated with precipitation in November ($R^2 = -0.339$, $P < 0.001$).

Discussion

What maintains the variation in complex traits in loblolly pine?

Dissecting the genetic and genomic architecture of complex traits is essential to understand the forces maintaining the variation in phenotypic traits of ecological and economical importance. In this study, we used newly generated genomic resources to gain insights into the genomic architecture of a large number of complex traits in a nonmodel plant species. Our results suggest a polygenic basis of quantitative inheritance, with significant differences in the number, effect size, genomic location and frequency of alleles contributing to variation in phenotypes in the different traits studied. Our results also suggest that while mutation-selection balance might be shaping the genetic variation in metabolic traits, balancing selection is more likely to shape the variation in expression of xylem development genes.

Rare alleles contributing to metabolic trait variation

Perhaps most of the discussion regarding the evolutionary forces maintaining genetic variation in complex traits can also be understood as the relative roles of rare alleles (maintained by mutation) and common alleles (maintained by selection) in phenotypic variation. Detection of low-frequency alleles showing associations with complex traits is challenging because it requires large and genetically diverse populations. In GWAS studies, SNPs with low minor allele frequencies are usually discarded due to their potential confounding effects with genotyping errors (false-positive associations). However, recent studies have shown that rare alleles may play an important role in the genetic regulation of complex traits in plant species, and may even help explaining the 'missing heritability' in forest trees species (Fahrenkrog *et al.*, 2017). QTLs showing associations with rare alleles have also been previously found in loblolly pine (Eckert *et al.*, 2012; Lu *et al.*, 2017).

In our study, we found that rare alleles associated with metabolic QTLs, with maf equal or less than 0.02 have large effect sizes of up to 56%. Effect sizes and minor allele frequencies were negatively associated for the majority of metabolic traits studied, following a L-shaped distribution of effect sizes (Bost *et al.*, 1999). While most of the associated SNPs have small to moderate effect sizes, only a small number of low-frequency alleles had the higher effect sizes. This pattern is consistent with the one described by the mutation-selection balance theory, in which low-frequency alleles (mainly deleterious) will have the higher effects in the variation of complex traits. In this sense, mutations that affect a trait may be under selection mainly because they have pleiotropic effects on fitness. As a result, mutations with large effect will be kept at low frequency due to their deleterious effects (Eyre-Walker, 2010). In fact, it is being suggested that larger effect sizes might be correlated with low maf as a result of purifying selection across metabolite networks (Keightley, 1989). Evidence for an inverse correlation between maf and effect sizes in complex traits has also been suggested in maize (Wallace *et al.*, 2014), *Medicago truncatula* (Stanton-Geddes *et al.*, 2013) and *Capsella* sp. (Josephs *et al.*, 2015).

Mutation–selection balance assumes either a high mutation rate or if mutation rate at each locus is small, then a large number of loci is required per trait (Barton & Keightley, 2002). Our study shows that large numbers of loci contribute to metabolic trait variation in loblolly pine, reducing the need for high mutation rates, which are unlikely to occur in conifer species (De La Torre *et al.*, 2017). Plant metabolites, which are extremely diverse (> 200 000) have been subject to purifying and positive selection in species such as Arabidopsis and *Picea* (Benderoth *et al.*, 2006; Keeling *et al.*, 2010; Luo, 2015). In *Arabidopsis thaliana*, gene duplication, and neo-functionalization have driven the large

number of secondary metabolites present in the species (Benderoth *et al.*, 2006). Increased rates of gene duplication may also have contributed to the large diversity of secondary metabolites involved in chemical defense systems in *Picea glauca* and *Picea sitchensis* (Keeling *et al.*, 2010; Warren *et al.*, 2015).

Clustered distribution of QTLs

Several studies have found evidence that alleles contributing to adaptive trait variation are sometimes physically clustered together. Due to the decrease in fitness with increasing recombination rates, clusters of alleles tend to be located in genomic regions with low recombination (Yeaman, 2013). In our study, we found that, although SNPs associated with metabolic traits are widely dispersed in the genome of loblolly pine, there are distinctive clusters (metabolic hotspots) of associated SNPs in linkage groups 1, 3, 7, 9 and 12. All of the SNPs clustered in metabolic hotspots are within close proximity (< 4 cM apart), or sometimes are even co-located (zero recombination events) in the same linkage group. In *Arabidopsis thaliana*, metabolic hotspots were located in genomic regions previously identified as being subject to strong positive selection (selective sweeps; Chang *et al.*, 2010). Metabolic hotspots have also been suggested in rice (Chen *et al.*, 2014), but found to be absent in maize (Riedelsheimer *et al.*, 2012). To our knowledge, our study is the first to suggest hotspots for metabolite−genotypic associations in a nonmodel plant species.

We also find clusters in alleles associated with the expression of xylem development genes. For example, five SNPs associated with the expression of xyloglucan endotransglycosylase 3 (XET-3) are co-located in linkage group 12, suggesting strong linkage and zero recombination events among these alleles. Physical information shows that SNPs associated with trait XET-3 are located in three different scaffolds (scaffold 3992, 2046 and 1709), which, according to our results, are located in the same linkage group. In addition, five SNPs associated with horseradish peroxidase C2 (prxC-2) have been found in linkage group 1, and three SNPs associated with callose synthase 3 (Cas-3) have been found to be co-located in linkage group 5 and in the same scaffold. Finally, our study has been able to locate, for the first time for the species, a cluster of eight SNPs associated with pitch canker disease in linkage group 7.

Common alleles and environmental variation contribute to expression trait variation

In contrast with metabolic traits, in which low-frequency alleles have the higher effect sizes, the majority of SNPs associated with expression of xylem development genes showed, on average, higher minor allele frequencies associated with higher effect sizes.

In fact, the median for maf equals 0.2156, almost 10 times higher than the median maf for metabolic traits. A smaller number of associated SNPs per trait was also observed. The pattern observed, in which common alleles with intermediate frequencies are associated with heritable trait variation is consistent with a model of balancing selection (Barton & Keightley, 2002; Turelli & Barton, 2004). Under this model, natural selection might act directly or indirectly on the trait, maintaining variation at loci that have pleiotropic effects on the trait under study (Barton & Keightley, 2002). Common alleles occurring across the species' natural range also showed strong allelic effects and patterns of allele-specific expression variation in our study in loblolly pine. For enzymes xyloglucan endotransglycosylase 3 and horseradish peroxidase C2, genotypes showed variation in expression that is neither affected by population of origin nor by environmental differences (Fig. 4b,d). Evidence of within-population balancing selection or between-populations local adaptation, have come from GWAS studies conducted on rangewide samples of plant species. For example, a GWAS study in *A. thaliana* suggested that strong divergent selection maintains the variation in glucosinate composition across European populations (Brachi *et al.*, 2015).

A small group of expression traits was found to be associated with environmental variables such as precipitation, radiation and climate moisture deficit, suggesting some of the variation in expression traits occurs due to spatial and environmental heterogeneity in loblolly pine (Table S10). Most expression traits showing significant associations, which were mostly enzymes, were negatively correlated with precipitation during the summer and positively correlated with Hargreaves climatic moisture deficit. Water availability is the highest abiotic determinant of the survival and growth of loblolly pine (Eckert *et al.*, 2010a,b), therefore precipitation, radiation and moisture deficit are key environmental variables. For example, higher expression levels of MADS box protein, a transcription factor putatively acting as a heat shock protein binding were found with increased levels of climatic moisture deficit in May, and increased radiation in spring. Also, higher expression levels of xyloglucan endotransglycosylase 2 (XET-2), an enzyme involved in xylem development, were found with increased radiation during the spring and decreased precipitation and moisture deficit during the summer. This enzyme is also correlated with another eight expression traits including enzymes involved in lignin biosynthesis (laccase 3, laccase 7 and phenylalanine ammonia lyase-1), cell expansion (COBRA and KORRIGAN), as well as CTL1, importin and APL transcription factor. Unfortunately, from all the expression traits that correlated with environmental variables, we could only find one trait that was significantly associated with any of the SNPs under study. The expression of the enzyme myo-inositol-1-phosphate synthase was associated with SNP AX-172858860 located in linkage group 9, negatively associated with precipitation during August, and positively associated with moisture deficit during the summer. Whether the environment plays a role in heritable trait variation or just phenotype plasticity of expression of xylem development genes remains to be addressed in future studies of the species.

Polygenic basis for complex traits

Our study suggests a largely polygenic basis for quantitative trait variation and the presence of pleiotropic effects in loblolly pine.

This is coincident with previous studies in conifers, in which a very small number of traits with simple inheritance (generally disease resistance traits such white pine blister rust and fusiform rust) has been reported (Kinloch *et al.*, 1970; Wilcox *et al.*, 1996). In contrast with crop and model plant species' GWAS in which few large effect QTLs have been found, our study suggests the presence of a moderate to large numbers of QTLs with a majority of small to medium effects. Most crop species studied to date are self-pollinated, with slow LD decay and strong population structure. Our results are coincident with GWAS studies in humans, maize (most studied outcrossing plant) and other forest trees (Buckler *et al.*, 2009; Visscher *et al.*, 2012; Fahrenkrog *et al.*, 2017). In all of these, LD decays rapidly and population structure is weak to moderate.

When comparing the different types of traits studied, we found significant differences in the number of significant SNPs. For example, pitch canker resistance, carbon isotope and expression of xylem development genes were associated with 1–17 SNPs per trait; whereas metabolic traits were found to be associated with up to 448 SNPs. In the case of pitch canker disease resistance, our study reports 18 associated SNPs, twice or more the amount of previously reported associations for the same trait (Quesada *et al.*, 2010; Moraga-Suazo *et al.*, 2014; Lu *et al.*, 2017). In relation to carbon isotope discrimination, our study found only two associated SNPs, less than found in previous studies using the same phenotypic dataset (four SNPs in Lu *et al.*, 2017; and 14 SNPs in Cumbie *et al.*, 2011). Our study also reports a lower number of associations in expression of xylem development genes (54 associations, 54 SNPs and 13 traits) than previously reported in Cumbie *et al.* (2011) (88 associations, 80 SNPs and 32 traits). All of these associations, however are new for the species. Finally, in relation to metabolic traits, our study reveals a much larger number of single-locus associations (2254 SNPs) that those reported in Eckert *et al.* (2012) (28 SNPs), probably because of the much larger genomic coverage of our study.

The study of complex traits in long-generation tree species require genome-wide assessments in widely distributed natural populations. By using whole-genome resequencing data with the purpose of generating SNPs for GWAS analysis, our study provides the most complete assessment of genomic variation (coding and noncoding regions) in a conifer species to date, and generates a better understanding of the segregating SNP variation responsible for phenotypic variation. This GWAS study, as well as other recent genome-wide studies in the species (Lu *et al.*, 2017) have resulted in a wealth of genomic resources now available to practice genomic selection and marker-assisted breeding in loblolly pine.

Acknowledgements

This project was supported by the US Department of Agriculture/National Institute of Food and Agriculture ('PineRefSeq'; 2011-67009-30030) awarded to DBN at the University of California, Davis. The authors would like to thank Andrew Eckert for his help in selecting the individuals for the whole-genome resequencing analysis; Dana Nelson, Chuck Burdine and Patrick Cumbie for sample collection; Randi Famula for laboratory support; and students who helped with DNA extraction such as Thomas Cartwright and Annalisa Romero.

Author contributions

DBN and ARDLT designed the research; CHL, MWC and KS produced the resequencing data; DP and SLS performed the SNP calling analyses; ARDLT supervised all laboratory work, collected phenotypic data, performed all data analyses and wrote the manuscript; all authors approved the final manuscript.

ORCID

Marc W. Crepeau http://orcid.org/0000-0002-3639-3921
Amanda R. De La Torre http://orcid.org/0000-0001-6647-723X
Charles H. Langley http://orcid.org/0000-0001-6160-5503
Daniela Puiu http://orcid.org/0000-0002-2386-9265
Steven L. Salzberg http://orcid.org/0000-0002-8859-7432

References

Barton NH, Keightley PD. 2002. Understanding quantitative genetic variation. *Nature Reviews Genetics* **3**: 11–21.

Benderoth M, Textor S, Windsor AJ, Mitchell-Olds T, Gershenzon J, Kroymann J. 2006. Positive selection driving diversification in plant secondary metabolism. *Proceedings of the National Academy of Sciences, USA* **103**: 9118–9123.

Bost B, Dillmann C, de Vienne D. 1999. Fluxes and metabolic pools as model traits for quantitative genetics. I. The L-shaped distribution of gene effects. *Genetics* **153**: 2001–2012.

Brachi B, Meyer CG, Villoutreix R, Platt A, Morton TC, Roux F, Bergelson J. 2015. Coselected genes determine adaptive variation in herbivore resistance throughout the native range of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* **112**: 4032–4037.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC *et al.* 2009. The genetic architecture of maize flowering time. *Science* **325**: 714–718.

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**: 7.

Chang EKF, Rowe HC, Hansen BG, Kliebenstein DJ. 2010. The complex genetic architecture of the metabolome. *PLoS Genetics* **6**: e1001198.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2**: 379–384.

Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, Li Y, Liu X, Zhang H, Dong H *et al.* 2014. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature Genetics* **46**: 714–721.

Cumbie WP, Eckert A, Wegrzyn J, Whetten R, Neale D, Goldfarb B. 2011. Association genetics of carbon isotope discrimination, height and foliar nitrogen in a natural population of *Pinus taeda* L. *Heredity* **107**: 106–114.

De La Torre AR, Birol I, Bousquet J, Ingvarsson PK, Jansson S, Jones SJM, Keeling CI, MacKay J, Nilsson O, Ritland K *et al.* 2014. Insights into conifer giga-genomes. *Plant Physiology* **166**: 1–9.

De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK. 2017. Contrasting rates of molecular evolution and patterns of selection among gymnosperms and flowering plants. *Molecular Biology and Evolution* **34**: 1363–1377.

Eckert AJ, Bower AD, Gonzalez-Martinez SC, Wegrzyn JL, Coop G, Neale DB. 2010a. Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology* **19**: 3789–3805.

Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, Gonzalez-Martinez SC, Neale DB. 2010b. Patterns of population structure and environmental association to aridity across the range of loblolly pine (*Pinus taeda* L, Pinaceae). *Genetics* **185**: 962–982.

Eckert AJ, Wegrzyn JL, Cumbie WP, Goldfarb B, Huber DA, Tolstikov V, Fiehn O, Neale DB. 2012. Association genetics of the loblolly pine (*Pinus taeda*, Pinaceae) metabolome. *New Phytologist* **193**: 890–902.

Eyre-Walker A. 2010. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences, USA* **107**: 1752–1756.

Fahrenkrog AM, Neves LG, Resende MFR, Vasquez AI, de los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB *et al.* 2017. Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytologist* **213**: 799–811.

Fuentes-Utrilla P, Goswami C, Cottrell JE, Pong-Wong R, Law A, A'Hara SW, Lee SJ, Woolliams JA. 2017. QTL analysis and genomic selection using RADseq derived markers in Sitka spruce: the potential utility of within family data. *Tree Genetics & Genomes* **13**: 33.

Gonzalez-Martinez SC, Huber D, Ersoz E, Davis JM, Neale DB. 2008. Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. *Heredity* **101**: 19–26.

Gonzalez-Martinez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB. 2007. Association genetics in *Pinus taeda* L. I. Wood properties traits. *Genetics* **175**: 399–409.

Groover A, Devey M, Fiddler T, Lee J, Megraw R, Mitchel-Olds T, Sherman B, Vujcic S, Williams C, Neale D. 1994. Identification of quantitative trait loci influencing wood specific gravity in an outbred pedigree of loblolly pine. *Genetics* **138**: 1293–1300.

Hall D, Hallingbäck HR, Wu HX. 2016. Estimation of number and size of QTL effects in forest tree traits. *Tree Genetics & Genomes* **12**: 110.

Jombart T. 2008. Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**: 1403–1405.

Jombart T, Ahmed I. 2011. Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**: 3070–3071.

Josephs EB, Lee YW, Stinchcombe JR, Wright SI. 2015. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proceedings of the National Academy of Sciences, USA* **112**: 15390–15395.

Josephs EB, Stinchcombe JR, Wright SI. 2017. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytologist* **214**: 21–33.

Keeling CI, Dullat HK, Yuen M, Ralph SG, Jancsik S, Bohlmann J. 2010. Identification and functional characterization of monofunctional ent-copalyl diphosphate and ent-kaurene synthases in white spruce reveal different patterns for diterpene synthase evolution for primary and secondary metabolism in gymnosperms. *Plant Physiology* **152**: 1197–1208.

Keightley PD. 1989. Models of quantitative variation of flux in metabolic pathways. *Genetics* **121**: 869–876.

Kinloch B, Parks GK, Fowler CW. 1970. White pine blister rust: simply inherited resistance in sugar pine. *Science* **167**: 193–195.

Lande R. 1975. The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genetic Research* **26**: 221–235.

Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie2. *Nature Methods* **9**: 357–359.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**: 2397–2399.

Lu M, Krutovsky KV, Nelson CD, West JB, Reilly NA, Loopstra CA. 2017. Association genetics of growth and adaptive traits in loblolly pine (*Pinus taeda* L.) using whole-exome-discovered polymorphisms. *Tree Genetics & Genomes* **13**: 57.

Luo J. 2015. Metabolite-based genome-wide association studies in plants. *Current Opinion in Plant Biology* **24**: 31–38.

Mitchell-Olds T, Willis JH, Goldstein DB. 2007. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Genetics* **8**: 845–856.

Moraga-Suazo P, Orellana L, Quiroga P, Balocchi C, Sanfuentes E, Whetten RW, Hasbun R, Valenzuela S. 2014. Development of a genetic linkage map for *Pinus radiata* and detection of pitch canker disease resistance associated QTLs. *Trees* **28**: 1823–1835.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* **12**: 111–121.

Neale DB, Martinez-Garcia PJ, De La Torre AR, Montanari S, Wei X. 2017. Novel insights into tree biology and genome evolution as revealed through genomics. *Annual Review of Plant Biology* **68**: 13.1–13.27.

Neale DB, Savolainen O. 2004. Association genetics of complex traits in conifers. *Trends in Plant Science* **9**: 325–330.

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD *et al.* 2014. Decoding the massive genomes of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology* **15**: R59.

Palle SR, Seeve CM, Eckert AJ, Cumbie WP, Goldfarb B, Loopstra C. 2011. Natural variation in expression of genes involved in xylem development in loblolly pine (*Pinus taeda* L.). *Tree Genetics & Genomes* **7**: 193–206.

Palle SR, Seeve CM, Eckert AJ, Wegrzyn JL, Neale DB, Loopstra C. 2013. Association of loblolly pine xylem development gene expression with single nucleotide polymorphisms. *Tree Physiology* **33**: 763–774.

Plomion C, Bastien C, Bogeat-Triboulot MB, Bouffier L, Dejardin A, Duplessis S, Fady B, Heuertz M, Le Gac AL, Le Provost G *et al.* 2016. Forest tree genomics: 10 achievements from the past 10 years and future prospects. *Annals of Forest Science* **73**: 77–103.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly MJ *et al.* 2007. PLINK: a tool set for whole-genome and population-based linkage analyses. *American Journal of Human Genetics* **81**: 559–575.

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

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altman T, Stitt M, Willmitzer L, Melchinger AE. 2012. Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proceedings of the National Academy of Sciences, USA* **109**: 23.

Stanton-Geddes J, Paape T, Epstein B, Briskine R, Yoder J, Mudge J, Bharti AK, Farmer AD, Zhou P, Denny R *et al.* 2013. Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*. *PLoS ONE* **8**: e65688.

Turelli M, Barton NH. 2004. Polygenic variation maintained by balancing selection: pleiotropy, sex-dependent allelic effects and GxE interactions. *Genetics* **166**: 1053–1079.

Turner SD. 2014. Qqman: an R package for visualizing GWAS results using Q-Q and Manhattan plots. *bioRxiv*. doi: 10.1101/005165.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *American Journal of Human Genetics* **90**: 7–24.

Wallace JG, Bradbury PJ, Zhang N, Gibon Y, Stitt M, Buckler ES. 2014. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genetics* **10**: e1004845.

Wang T, Hamann A, Spittlehouse D, Carroll C. 2016. Locally downscaled and spatially customizable climate data for historical and future periods for North America. *PLoS ONE* **11**: e0156720.

Warren RL, Keeling CI, Yuen MMS, Raymond A, Taylor GA, Vandervalk BP, Mohamadi H, Paulino D, Chiu R, Jackman SD *et al.* 2015. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families

of conifer terpenoid and phenolic defense metabolism. *The Plant Journal* **83**: 189–212.

Wilcox PL, Amerson HV, Kuhlman EG, Liu BH, O'Malley DM, Sederoff RR. 1996. Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. *Proceedings of the National Academy of Sciences, USA* **93**: 3859–3864.

Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adapted loci. *Proceedings of the National Academy of Sciences, USA* **110**: e1743–e1751.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM *et al.* 2010. Mixed linear model approach adapted for genome wide association studies. *Nature Genetics* **42**: 355–360.

Zheng Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7**: 203–214.

Zimin AV, Stevens KA, Crepeau MW, Puiu D, Wegrzyn JL, Yorke JA, Langley CH, Neale DB, Salzberg SL. 2017. An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *GigaScience* **6**: 1–4.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Relationship between effect size and minor allele frequency.

Methods S1 SNP calling.

Table S1 List of significant phenotypes.

Table S2 Results of the GWAS analysis for morphological traits: pitch canker disease resistance and carbon isotope discrimination.

Table S3 Genomic location and annotation of mQTLs located in hotspots.

Table S4 Results of the GWAS analysis for molecular traits: metabolites.

Table S5 Results of the GWAS analysis for molecular traits: expression of xylem development genes.

Table S6 Allelic effects for pitch canker disease resistance.

Table S7 Allelic effects for metabolites.

Table S8 Allelic effects for expression of xylem development genes.

Table S9 Correlations between minor allele frequency and effect size.

Table S10 Correlations between phenotypes and environmental variables.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

See also the Commentary on this article by Casola, **221**: 1669–1671.