










# Providing biological context for GWAS results using eQTL regulatory and co-expression networks in *Populus*

Mengjun Shu<sup>1,2</sup> , Timothy B. Yates<sup>1,2</sup>, Cai John<sup>1,2,3</sup>, Anne E. Harman-Ware<sup>4</sup> , Renee M. Happs<sup>4</sup> , Nathan Bryant<sup>3</sup> , Sara S. Jawdy<sup>1,2</sup> , Arthur J. Ragauskas<sup>1,2,3</sup> , Gerald A. Tuskan<sup>1,2</sup> , Wellington Muchero<sup>1,2†</sup>  and Jin-Gui Chen<sup>1,2</sup> 

<sup>1</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge 37831 TN, USA; <sup>2</sup>Center for Bioenergy Innovation, Oak Ridge National Laboratory, Oak Ridge 37831 TN, USA;

<sup>3</sup>Department of Chemical and Biomolecular Engineering, University of Tennessee, Knoxville 37996 TN, USA; <sup>4</sup>Renewable Resources and Enabling Sciences Center, National Renewable Energy Laboratory, Golden 80401 CO, USA

## Summary

Author for correspondence:  
Jin-Gui Chen  
Email: [chenj@ornl.gov](mailto:chenj@ornl.gov)

Received: 19 February 2024  
Accepted: 16 July 2024

New Phytologist (2024) 244: 603–617  
doi: 10.1111/nph.20026

**Key words:** expression quantitative trait loci, gene networks, genome-wide association studies, lignocellulose formation, *Populus*.

- Our study utilized genome-wide association studies (GWAS) to link nucleotide variants to traits in *Populus trichocarpa*, a species with rapid linkage disequilibrium decay. The aim was to overcome the challenge of interpreting statistical associations at individual loci without sufficient biological context, which often leads to reliance solely on gene annotations from unrelated model organisms.
- We employed an integrative approach that included GWAS targeting multiple traits using three individual techniques for lignocellulose phenotyping, expression quantitative trait loci (eQTL) analysis to construct transcriptional regulatory networks around each candidate locus and co-expression analysis to provide biological context for these networks, using lignocellulose biosynthesis in *Populus trichocarpa* as a case study.
- The research identified three candidate genes potentially involved in lignocellulose formation, including one previously recognized gene (Potri.005G116800/VND1, a critical regulator of secondary cell wall formation) and two genes (Potri.012G130000/AtSAP9 and Potri.004G202900/BIC1) with newly identified putative roles in lignocellulose biosynthesis.
- Our integrative approach offers a framework for providing biological context to loci associated with trait variation, facilitating the discovery of new genes and regulatory networks.

## Introduction

Population genomics enables the identification of genetic loci underlying spatial and temporal patterns of phenotypic variation, a long-standing goal of evolutionary biology (Feder & Mitchell-Olds, 2003; Weigel & Nordborg, 2015). *Populus* species (poplars, aspens, and cottonwoods) are excellent models for population genomic studies of woody plants because of their modest genome size (*c.* 480 Mbp), dioecious nature, relative ease of transgenic manipulation, rapid growth, broad geographic

Notice: This manuscript has been authored by UT-Battelle, LLC under contract no.: DE-AC05-00OR22725 with the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

†Deceased.

distribution, and rich genetic resources (Taylor, 2002; Wullschlegel *et al.*, 2002; Davis *et al.*, 2006; Tuskan *et al.*, 2006; Rubin, 2008). *Populus* species and hybrids are widely regarded as leading bioenergy production feedstock for conversion to biofuels (Sannigrahi *et al.*, 2010; Studer *et al.*, 2011; Porth & El-Kassaby, 2015). The plant secondary cell wall (SCW) provides the source of lignocellulosic biomass, composed of cellulose, hemicellulose, and lignin (Kumar *et al.*, 2016; Meents *et al.*, 2018). Many studies have explored various aspects related to the development of lignocellulosic biosynthesis and deconstruction in *Populus*, such as the saccharification of cellulose and hemicellulose to simple sugars via hydrolysis, the effects of lignin content and composition on sugar release, and the fermentation of sugars to ethanol (Groover *et al.*, 2010; Studer *et al.*, 2011; Welker *et al.*, 2015; Marriott *et al.*, 2016). Similarly, significant research has uncovered genes and genetic regulation mechanisms of complex cell wall traits in *Populus* (Wegrzyn *et al.*, 2010; Guerra *et al.*, 2013, 2019; Zhang *et al.*, 2020; Ren *et al.*, 2022). Despite the progress, an integrative approach that combines genome-wide association studies (GWAS), expression quantitative trait loci (eQTL) analyses, and co-expression network

analyses to unravel the regulatory networks of lignocellulose phenotypes remains underexplored. Our study addresses this gap by applying a multifaceted analytical approach to reveal the genetic architecture of lignocellulosic biomass development in *Populus*.

Association genetics approaches, such as GWAS, have emerged as powerful tools for identifying the genomic regions associated with complex traits in plants (Syvänen, 2005; Ingvarsson & Street, 2011). GWAS has generated robust associations for a range of plant traits, including model and nonmodel systems (Alseekh *et al.*, 2021; Tibbs Cortes *et al.*, 2021; Demirjian *et al.*, 2023). Despite its strengths, GWAS is inherently limited by the statistical power, which depends on the number of loci and individuals analyzed, affecting the capacity to detect associations between DNA variants and traits (Korte & Farlow, 2013; Visscher *et al.*, 2017). While high-throughput sequencing and genotyping have increasingly been utilized to address some of these challenges (Hamilton & Robin, 2012; D'Agostino & Tripodi, 2017; Bhat & Yu, 2021), a significant enhancement in GWAS's power is achieved through the incorporation of a broader spectrum of phenotypic measurements obtained via various technical approaches. This extension within the univariate GWAS framework enhances the detection capabilities of genetic associations by leveraging the complexity of the traits, thereby reducing false positives and significantly improving the discovery of genetic variants (Ritchie *et al.*, 2015; Porter & O'Reilly, 2017; Chhetri *et al.*, 2019).

Another typical limitation of GWAS is the difficulty in interpreting the results biologically (Visscher *et al.*, 2017). eQTL analysis, which maps quantitatively measured gene expression variation to a genomic locus, can partially address this limitation by identifying local *cis* and remote *trans* genetic elements that regulate the expression levels of critical genes correlated with traits of interest (Zhu *et al.*, 2016; Li *et al.*, 2020). Integrating GWAS and eQTL datasets facilitates the biological interpretation of GWAS findings (Gamazon *et al.*, 2015). For example, studies have successfully combined GWAS with eQTL analysis in *Populus* to identify genes determining leaf shape characteristics (Mähler *et al.*, 2020) and regulators for secondary metabolite biosynthesis (Zhang *et al.*, 2018b). Moreover, this integration increases statistical power by focusing on the genetic component of expression, excluding environmental factors influencing gene expression and complex traits (Xu *et al.*, 2017). Despite the advantages of combining GWAS and eQTL, and the importance of lignocellulosic biomass, few studies have applied these methods in *Populus* to dissect the genetic architecture of regulatory variation underlying traits related to lignocellulose.

In this study, we analyzed transcriptomic and phenotypic data from black cottonwood (*Populus trichocarpa* Torr. & A. Gray) by integrating GWAS and eQTL analyses to identify single-nucleotide polymorphisms (SNPs) and gene networks impacting lignocellulosic cell wall biosynthesis. For the GWAS analysis, we measured multiple lignocellulose-related traits using several independent techniques, including pyrolysis molecular beam mass spectrometry (py-MBMS), proton nuclear magnetic resonance ( $^1\text{H}$  NMR) analysis, and heteronuclear single quantum coherence (HSQC) NMR spectroscopy. These methods provided an estimate of five-carbon hemicellulose sugar content, six-carbon

cellulose sugar content, lignin content, and monolignol ratio in our association population. After characterizing candidate genes associated with lignocellulose in the GWAS, we conducted eQTL mapping to identify putative regulatory networks encompassing these genes. Furthermore, we performed co-expression analysis to provide additional evidence supporting the importance of the candidate genes. Collectively, these analyses provided insights into the regulatory network involving a set of genes implicated in the chemical wood properties of *P. trichocarpa*.

## Materials and Methods

### Lignocellulose phenotyping in the Corvallis black cottonwood collection

Wood samples were harvested from 834 genotypes of 3-yr-old *Populus trichocarpa* Torr. & A. Gray grown in a common garden in Corvallis, OR (44°34'14.81"N 123°16'33.59"W). The sampling process followed previously described methods (Muchero *et al.*, 2015; Chhetri *et al.*, 2019). In brief, increment cores with a diameter of 1 cm were extracted at breast height (1.3 m) from each tree using an increment borer. The collected cores were immediately placed in zip-lock bags and stored at  $-20^\circ\text{C}$  to prevent any degradation or compositional changes before processing. Before analysis, the wood cores were air-dried at room temperature. Once dried, they were debarked. The debarked wood samples were then ground into a fine powder using a Wiley mill with an 80/20 mesh. We collected and measured two distinct sets of samples for lignin-related and carbohydrate phenotypes using the various analytical methods listed above.

The first set consisted of 404 individual samples analyzed using HSQC NMR spectroscopy to evaluate the lignin-related phenotypes. This methodology and the associated measurements have been detailed previously in Bryant *et al.* (2023).

The second set included a larger collection of 749 samples, with 319 individuals coinciding with the first set. We measured the phenotype for the second set using two methods, including py-MBMS and  $^1\text{H}$  NMR. For py-MBMS, we calculated the lignin-related phenotype based on the *m/z* values from *m/z* 30 to *m/z* 450, resulting in estimates of lignin content and the S/G ratio, where S represents syringyl forms of lignin and G represents guaiacyl forms of lignin. py-MBMS was performed as described previously (Harman-Ware *et al.*, 2022). Briefly, 4 mg of debarked, milled, enzymatically destarched, and ethanol-extracted biomass samples were pyrolyzed for 30 s at  $500^\circ\text{C}$  under a helium atmosphere using a Frontier Py2020 pyrolyzer interfaced to an Extrel Max 1000 Molecular Beam Mass Spectrometer. The spectrometer ionization was set to  $-17$  eV and recorded spectra in centroid mode from 30 to 450 *m/z*.

For  $^1\text{H}$  NMR, we measured the cell wall carbohydrate phenotype from NMR analysis of two-stage acid hydrolysates, focusing on glucose, xylose, galactose, arabinose, and mannose. This method was performed as described previously (Happs *et al.*, 2021). Briefly,  $^1\text{H}$  spectra of hydrolysates were collected at  $25^\circ\text{C}$  with a Bruker 5 mm BBO probe using NOESY 1D presaturation, 64 scans, and a 5 s recycle delay. All spectra were

processed in Topspin 3.5pl7 using standard parameters. Bruker's AMIX software was used to divide spectra into 0.005 ppm buckets in the region of 3.10–4.15 ppm. PLS models for five monomeric sugars were built using HPLC-derived sugar concentrations from hydrolysates of a standard sample set and were performed in the UNSCRAMBLER v.10.5 (CAMO A/S, Trondheim, Norway).

In total, across all platforms, we measured 47 phenotypes (four different sets) across 834 genotypes, as documented in Supporting Information Tables S1–S3 and Fig. S1. This dataset included a list of 28 *m/z* values measured by py-MBMS that has been assigned to carbohydrate or lignin content in previous studies (749 samples, py-MBMS\_ *m/z*), two lignin-related measurements calculated by *m/z* values in py-MBMS (749 samples, py-MBMS\_lignin), six carbohydrate phenotypes measured by <sup>1</sup>H NMR (749 samples, NMR\_C5C6, where C5 represents five-carbon carbohydrates, and C6 represents a six-carbon carbohydrate, glucose), and 11 lignin measurements by HSQC NMR (404 samples, HSQC\_NMR\_lignin). Pearson's correlations among phenotypic traits were investigated and visualized (Fig. S2) using R software package 'CORRPLOT' (Wei & Simko, 2017). The statistical significance of the correlations was assessed at a confidence level of 95%, corresponding to a *P*-value threshold of 0.05, to ensure robustness.

### Genome-wide association mapping

Whole-genome DNA short-read sequencing was performed on 1323 *P. trichocarpa* genotypes, including 834 assessed for lignocellulose phenotyping (Tables S1–S3), using Illumina Genome Analyzer, HiSeq 2000, and HiSeq 2500 platforms as per protocols described by Evans *et al.* (2014). This approach ensured a minimum expected sequencing depth of 15×. The sequence data were subjected to SNP calling (see Methods S1 for detailed protocol). In total, 9751 445 SNPs and indel variants with minor allele frequency (MAF) > 0.05 were identified. This SNP dataset is available at doi: 10.25983/2352478.

For GWAS analyses, SNPs corresponding to the genotypes used for the various lignocellulose measurements (detailed in Tables S2, S3) were analyzed. Approximately 8.3 million SNPs were evaluated for each of the 47 phenotypic traits using a univariate linear mixed model implemented in GEMMA (Zhou & Stephens, 2012). The built-in estimation of a centered relatedness matrix to control population structure in GEMMA software was used as the correction factor for genetic background effects. Deviation of *P*-values from normality was assessed using quantile–quantile (Q–Q) plots (Fig. S3).

We established a stringent *P*-value threshold of  $P < 10^{-7}$  for declaring significant SNP–trait associations, based on the Bonferroni correction method for multiple testing (effective threshold of  $P < 0.05/8300\ 000 = 10^{-8.2}$ ). This threshold ( $P < 10^{-7}$ ) was selected to balance the need for stringent filtering while accommodating the broad scope of phenotypic analyses. For significant SNP associations with phenotypic traits, we visualized the relationships using boxplots generated with the GGLOT2 package in R (Wickham, 2011). Before plotting, we normalized the measurements of the associated traits by applying the scale()

function in R, which standardizes each trait to have a zero mean and unit variance. This normalization process allows for a direct comparison across traits measured on different scales. For significant SNPs, we also performed *t*-test comparisons among genotypes for each associated phenotypic trait, assessing statistical significance at  $P < 0.05$  to ensure robustness.

For the same reference genome, *P. trichocarpa* v.3.0, two versions of the annotation file are available, including *P. trichocarpa* v.3.0 and v.3.1 ([https://phytozome-next.jgi.doe.gov/info/Ptrichocarpa\\_v3\\_0](https://phytozome-next.jgi.doe.gov/info/Ptrichocarpa_v3_0); [https://phytozome-next.jgi.doe.gov/info/Ptrichocarpa\\_v3\\_1](https://phytozome-next.jgi.doe.gov/info/Ptrichocarpa_v3_1)). We combined these two files before running annotation after GWAS. In total, 38 830 genes overlapped between these two annotation files. In the combined annotation file, the start of the overlapping gene was defined as the smaller one and the end as the larger one among the two versions to increase the chance of locating candidate genetic loci in genes. For the 2505 unique genes in *P. trichocarpa* v.3.0 and 4120 unique genes in v.3.1, we directly included the information in the new annotation file. Thus, the updated annotation file contains 45 455 genes. The combined annotation file was used to associate genes with SNPs, and any gene where a SNP, genic or intergenic, considered to be associated with that SNP. This combined annotation file is available at doi: 10.25983/2352478.

### RNA sequencing and eQTL analysis

RNA extraction, library construction, and sequencing methodologies were conducted according to the protocols described by Zhang *et al.* (2018b). Sequencing was performed on the Illumina HiSeq 2500 platform using a paired-end 150-bp configuration. For the eQTL analysis, we used RNA-Seq data derived from xylem tissue in 533 genotypes from our 1323 GWAS population. Detailed information about this dataset, including NCBI SRA accession numbers for the publicly available sequences, is found in Tables S3 and S4.

The analytical approach for eQTL analysis followed the approach outlined by Yates *et al.* (2021). Briefly, raw RNA-Seq reads across all genotypes underwent quality control and trimming using the JGI QC pipeline, which employs BBDuk (<https://sourceforge.net/projects/bbmap/>). Next, we aligned the processed reads to the *P. trichocarpa* v.3.1 reference genome using STAR v.2.6.1b. Gene count quantification was performed with FEATURECOUNTS v.1.6.3, and counts were normalized to counts per million (CPM) using EDGER (Robinson *et al.*, 2010; Dobin *et al.*, 2013; Liao *et al.*, 2014).

For the eQTL analysis, we used GEMMA (Zhou & Stephens, 2012), similar to our GWAS methodology, ensuring control for population structure. The built-in estimation of a centered relatedness matrix in GEMMA was used as the correction factor for genetic background effects. Instead of using phenotypic measurements as input, we used CPM per gene, ensuring a direct link between gene expression profiles and genetic variation. The SNPs were sourced exclusively from those genotypes for which we had corresponding expression data, *c.* 7.8 million.

All genes with expression evidence in xylem tissue, totaling 39 380, were used in the eQTL analysis. We established

a stringent  $P$ -value threshold of  $P < 10^{-7}$  for declaring significant eQTL, based on the Bonferroni correction method for multiple testing (effective threshold of  $P < 0.05/7800\ 000 = 10^{-8.2}$ ). Additionally, for an eQTL interval to be considered significant, at least five SNPs needed to be present in the peak. eQTLs located on different chromosomes than the target gene, or on the same chromosome but  $> 1$  Mb away from the target gene, were classified as *trans*-eQTLs. eQTLs on the same chromosome within 1 Mb of the target gene were classified as *cis*-eQTLs. In eQTL analysis, an expression quantitative trait nucleotide (eQTN) is defined as a specific SNP that is significantly associated with changes in gene expression level.

### Upstream and downstream gene set enrichment analysis

Downstream genes are defined as those whose expression levels are significantly associated with eQTNs identified in the eQTL analysis. Specifically, these downstream genes are putatively regulated by the eQTNs. For downstream analysis, we considered a 10-kb flanking interval around each eQTN to account for measurement-error-driven drift in association signals, thereby identifying the target genes.

Upstream analysis involves treating the candidate genes as target genes and identifying the eQTNs associated with their expression. Additionally, we identified all target genes associated with these upstream eQTNs within a flanking interval of 10 kb. This approach allowed us to map the putative genetic regulation network upstream of the candidate genes.

Based on the results of GWAS analysis, we defined the candidate SNPs as those significantly associated ( $P < 10^{-7}$ ) with at least two different sets of phenotypes. By mapping the candidate SNPs with the combined annotation file, the names of genes where each candidate SNP locate in or between were obtained. Using the *trans*-eQTL results, we identified downstream and upstream gene sets for all candidate SNPs and genes, providing a comprehensive view of the genetic regulation or causal network involved.

After obtaining sets of downstream genes that were associated with the candidate genes, as well as the upstream genes that were associated with the expression of the candidate genes, we performed Gene Ontology (GO) enrichment analysis following approaches reported by Subramanian *et al.* (2005). *Populus* genes were mapped to Arabidopsis TAIR IDs using the `bitr` function from the Bioconductor package `CLUSTERPROFILER` (Wu *et al.*, 2021), mapping gene IDs based on homology to the TAIR10. Enrichment analyses were conducted using the function `enrichGO` in Bioconductor packages `CLUSTERPROFILER` (Wu *et al.*, 2021) with the database `org.At.tair.db` (Carlson, 2021). R function `enrichGO` automates the process of biological-term classification and the enrichment analysis of gene clusters. We applied the false discovery rate (FDR) correction using the Benjamini–Hochberg procedure to control for multiple testing at an FDR threshold of 0.05. The output was visualized by bar plot, enrichment map, and category-gene network plot using the R package `enrichplot` (Yu, 2022).

### Co-expressed genes analysis

To construct co-expression networks for candidate genes, we employed the weighted gene co-expression network analysis (WGCNA) approach in R (Langfelder & Horvath, 2008), using RNA sequencing data from xylem tissue. We calculated the Pearson correlation coefficients for all gene pairs and transformed these into an adjacency matrix using a soft-thresholding power of seven to maintain scale-free topology. The network constructed was unsigned to capture both positive and negative correlations.

We then calculated the topological overlap matrix (TOM) from the adjacency matrix and derived the dissimilarity TOM (dissTOM) for hierarchical clustering. Genes were clustered hierarchically based on the dissTOM, and the dynamic tree-cut algorithm was applied for module detection, with a deep split of 2 and a minimum module size of 30 genes. Each module was assigned a unique color for visualization purposes. We used TOM to identify the most strongly interconnected genes within each module, thereby reinforcing the biological relevance of our co-expression networks. Upon obtaining sets of co-expressed genes for our candidate gene, we carried out GO enrichment analysis following the same methodology as for the downstream and upstream genes.

## Results

### Phenotyping and GWAS analyses

To characterize the chemical properties of the cell wall in the *P. trichocarpa* natural variants, we measured 47 lignocellulose traits across 834 genotypes using three different methods: py-MBMS ( $m/z$  & lignin composition), HSQC NMR (lignin linkages) of extracted lignin, and  $^1\text{H}$  NMR of two-stage acid hydrolysates (carbohydrate). Briefly, we examined 28  $m/z$  values determined by py-MBMS and assigned to carbohydrate or lignin content (Harman-Ware *et al.*, 2021), two lignin-related measurements calculated by  $m/z$  values in py-MBMS, six carbohydrate phenotypes measured by  $^1\text{H}$  NMR, and 11 lignin measurements estimated from HSQC NMR. The distribution and correlation patterns of these traits are illustrated in Figs S1 and S2, respectively, highlighting the extensive phenotypic variation and inter-relationships among the measured traits.

GWAS analyses results for the lignin and carbohydrate traits revealed a large set of significantly associated SNPs ( $P < 10^{-7}$ ) corresponding to each phenotype. Specifically, for the  $m/z$  values measured by py-MBMS (749 samples), we identified 91 significant genes associated with 99 SNP loci. In the case of lignin-related measurements calculated by  $m/z$  values in py-MBMS (749 samples), we found four significant genes associated with two SNP loci. For the carbohydrate phenotypes measured by  $^1\text{H}$  NMR (749 samples), we detected 13 significant genes associated with 10 SNP loci. Finally, for the lignin linkage measurements by HSQC NMR (411 samples), we identified 490 significant genes associated with 588 SNP loci. To identify a shared set of genes associated with lignocellulose traits in *P. trichocarpa*, we examined the overlap of significantly associated SNPs across the

analytical platforms. Although most associations were specific to individual traits (Table S5), we detected 13 genes with associations in two or more GWAS analyses (Fig. 1; Table S6).

### Candidate gene across py-MBMS\_ *m/z*, py-MBMS\_lignin and NMR\_C5C6

Among all significant associations, SNP position Chr12:14 692 802 was notably associated with three distinct phenotypic datasets (Table S2). These include nine traits from py-MBMS\_ *m/z* group (*m/z*: 57, 60, 73, 98, 126, 147, 144, 154, 180), one trait from the py-MBMS\_lignin group, and one trait from the NMR\_C5C6 group related to the glucose–xylose ratio (Table S6; Fig. 2a). Using *m/z* 180 as a representative peak, a Manhattan plot was constructed for the py-MBMS\_ *m/z* group, complemented by traits from the other two groups. Notably, this SNP, Chr12:14 692 802, exhibited the lowest *P*-value on Chromosome 12 among the traits from the three phenotyping platforms (Fig. 2a–c).

We normalized the measurements of the 11 associated traits noted above to have zero mean and unit variance and constructed boxplots for each trait across different genotypes (Fig. 2d). For the glucose–xylose ratio in NMR\_C5C6 and six py-MBMS\_ *m/z* traits (*m/z*: 57, 60, 73, 98, 126, and 144), a consistent pattern emerged: The homozygous SNP genotype (TT) showed lower measurements than the heterozygous SNP genotype (T/C). By contrast, three other py-MBMS\_ *m/z* traits (*m/z*: 137, 154, and 180) and one py-MBMS\_lignin trait (lignin content) displayed higher measurements in the TT genotype.

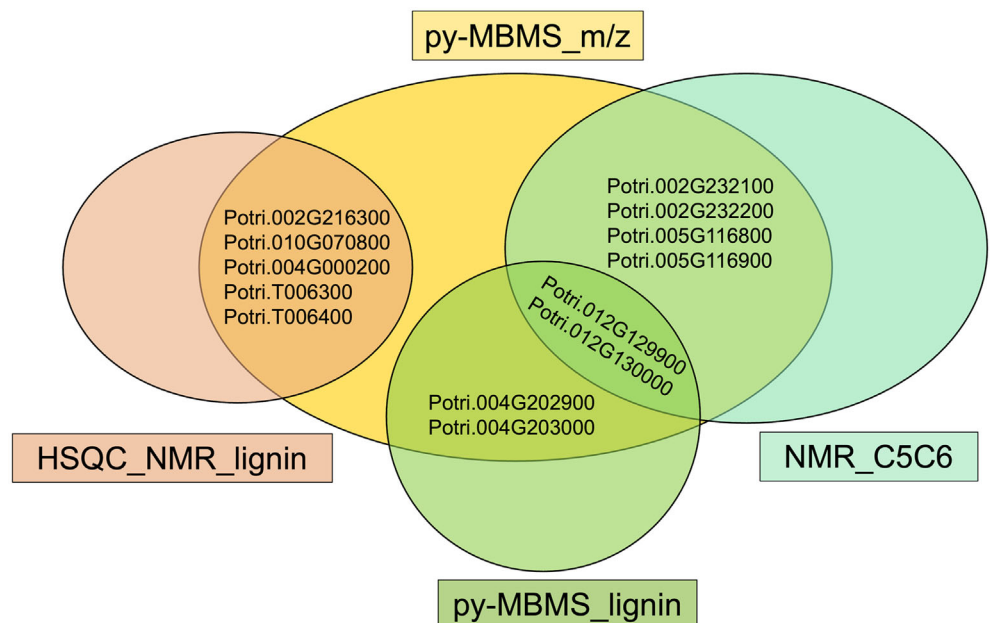
Correlation analyses highlighted two distinct groups with significant intra-group positive correlations (Fig. 2e). The first group, comprising the glucose–xylose ratio and six py-MBMS\_ *m/z* traits (*m/z*: 57, 60, 73, 98, 126, and 144), showed strong positive correlations. The second group, three py-MBMS\_ *m/z*

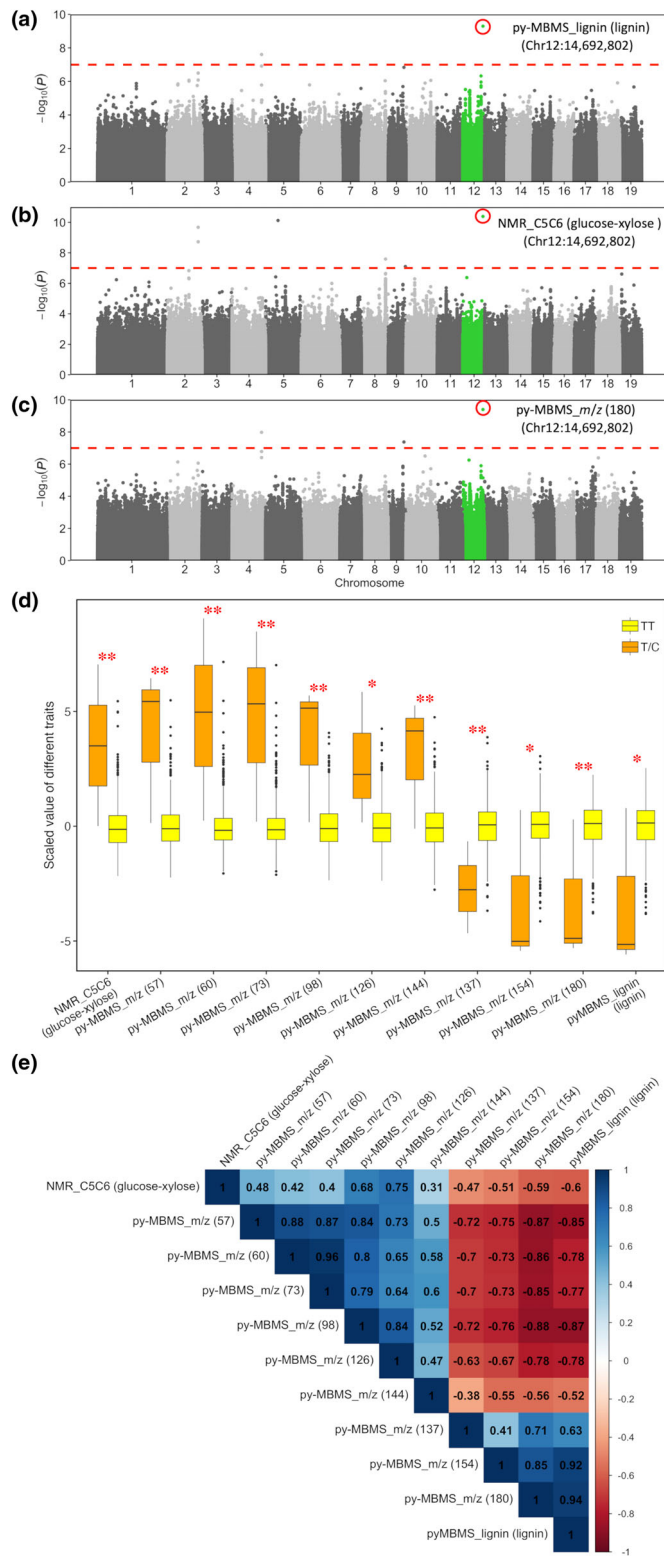
traits (*m/z*: 137, 154, and 180) and one py-MBMS\_lignin trait (lignin content), also exhibited strong positive correlations. Conversely, these two groups demonstrated significant negative inter-group correlations.

Previous studies have linked specific pyrolysis products to define py-MBMS *m/z* values as indicators of carbohydrate and lignin components in *Populus* wood samples (Sykes *et al.*, 2008; Xiao *et al.*, 2014; Harman-Ware *et al.*, 2021, 2022). For instance, *m/z* values 57 (C5, C6), 60 (C5, C6), 73 (C5, C6), 98 (C6), 126 (C6), and 144 (C6) have been designated to C5 and C6 sugars. Additionally, *m/z* values 137 (G), 154 (S), and 180 (S, G) have been shown to be associated with lignin, representing syringyl (S) and guaiacyl (G) components (Sykes *et al.*, 2008; Xiao *et al.*, 2014; Harman-Ware *et al.*, 2021, 2022). Our findings corroborate these reports, with patterns observed in *m/z* 57, 60, 73, 98, 126, and 144 aligning with carbohydrate traits, and *m/z* 137, 154, and 180 reflecting lignin components in both the boxplots and the correlation plot (Fig. 2d,e). The divergence between carbohydrate-related and lignin-related traits suggests a metabolic trade-off between structural constituents and intracellular chemical components in *P. trichocarpa* (Harman-Ware *et al.*, 2022).

In addition to GWAS analyses, we incorporated *trans*-eQTL approaches to identify downstream genes associated with SNP Chr12:14 692 802 in xylem tissue. We conducted a comprehensive eQTL analysis across the genome, focusing on eQTNs with *P*-values  $<10^{-7}$ . By examining the eQTL results of downstream genes associated with SNP Chr12:14 692 802, we conducted a GO enrichment analysis to understand the biological functions of these downstream genes. The predominant GO terms linked to these downstream genes were associated with lignin biosynthesis, including phenylpropanoid biosynthetic process (GO:0009699), phenylpropanoid metabolic process (GO:0009698), secondary metabolite biosynthetic process

**Fig. 1** Overlapping genes identified via genome-wide association studies (GWAS) analyses on four sets of lignocellulose traits in *Populus trichocarpa*. Traits include *m/z* values measured by pyrolysis molecular beam mass spectrometry (py-MBMS) assigned to carbohydrate or lignin content in previous studies (py-MBMS\_ *m/z*), lignin-related measurements calculated by *m/z* values in py-MBMS (py-MBMS\_lignin), carbohydrate phenotypes measured by proton nuclear magnetic resonance ( $^1\text{H}$  NMR) (NMR\_C5C6), and lignin measurement by heteronuclear single quantum coherence (HSQC) NMR (HSQC\_NMR\_lignin). C5 and C6 refer to five- and six-carbon carbohydrates, respectively.





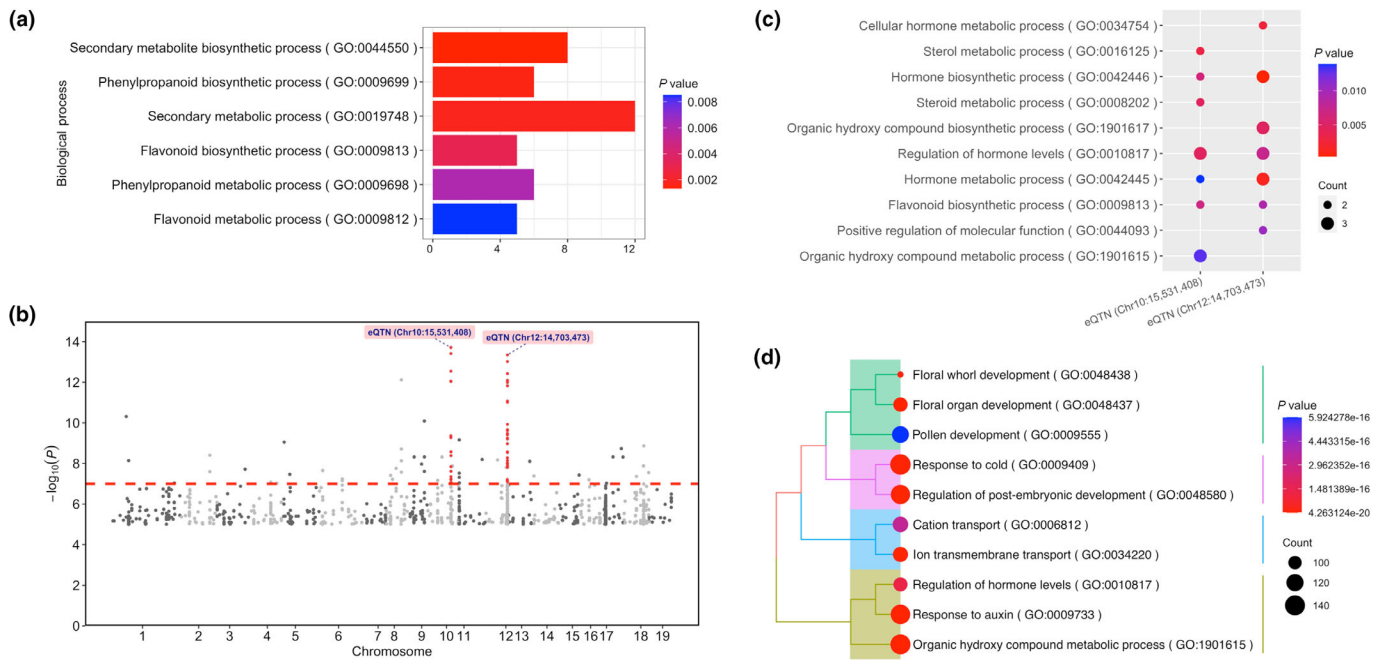
**Fig. 2** Manhattan plots and trait analyses for single nucleotide polymorphism (SNP) Chr12:14 692 802 in *Populus trichocarpa*. (a–c) Manhattan plots highlighting the genome-wide association studies (GWAS) signal at SNP Chr12:14 692 802 associated with py-MBMS\_lignin (lignin), NMR\_C5C6 (glucose-xylose), and py-MBMS\_m/z ( $m/z$ : 180). (d) Boxplot illustrating the 11 traits significantly associated with SNP Chr12:14 692 802 across different genotypes. *Note*: Traits measurements are normalized. Horizontal lines within each box represent the median, the boxes represent the interquartile range (IQR), the whiskers extend to 1.5 times the IQR, and dots represent outliers. The characters \* and \*\* indicate  $t$ -test  $P$ -values  $< 0.05$  and  $< 0.01$ , respectively. (e) Correlation plot of the same 11 traits showing Pearson correlation coefficients, with all correlations significant at a  $P$ -value threshold  $\leq 0.05$ .

Arabidopsis sequence homolog: AT5G48930: HCT), and Potri.006G024300 (closest Arabidopsis sequence homolog: AT1G72680: CAD1) have previously been linked to SCW regulation and biosynthesis (Li *et al.*, 2017; Eckert *et al.*, 2019; Su *et al.*, 2019). In summary, SNP Chr12:14 692 802 affects genes involved in lignin biosynthesis in *P. trichocarpa* xylem.

SNP Chr12:14 692 802 occurs between two annotated genes: Potri.012G129900 (closest Arabidopsis sequence homolog: AT5G51670) and Potri.012G130000 (closest Arabidopsis sequence homolog: AT4G22820: AtSAP9). To determine which of the two genes may be responsible for the association, we examined the network topology of both genes using eQTL data derived from xylem tissue expression. Only one of the two genes, Potri.012G130000, exhibited significant upstream eQTL nucleotides (eQTNs) ( $P < 10^{-7}$ ), including a *trans* upstream eQTN on Chromosome 10 and multiple *cis* regulators on Chromosome 12 (Fig. 3b). The target genes of these upstream eQTNs in xylem primarily function in hormone biosynthesis (GO:0034754, GO:0042446, GO:0010817, GO:0042445, and GO:1901615), as shown in Fig. 3(c). In the xylem tissues, hierarchical clustering revealed distinct modules of co-expressed genes, as visualized in the cluster dendrogram (Fig. S4). Notably, the xylem co-expression analysis of Potri.012G130000 suggested its involvement with genes related to hormone biosynthesis (GO:0010817, GO:0009733, and GO:1901615) (Fig. 3d). Plant hormones have been substantiated to influence cambial division, cell differentiation, and maturation, exerting a profound impact on wood quantity and quality in trees (Buttò *et al.*, 2020). Collectively, both eQTL and co-expression data point to Potri.012G130000 (AT4G22820: AtSAP9) as the most likely candidate gene linked to the SNP at Chr12:14 692 802.

Potri.012G130000 (AT4G22820: AtSAP9) is annotated as ZINC FINGER A20 AND AN1 DOMAIN-CONTAINING STRESS-ASSOCIATED PROTEIN 10-RELATED, and exhibited significant associations in our GWAS analyses with py-MBMS\_m/z, py-MBMS\_lignin, and NMR\_C5C6 phenotypes. Moreover, we delineated both upstream and downstream gene networks for this SNP based on our *trans*-eQTL results. In the context of xylem tissue, downstream genes were associated with lignin biosynthesis and the upstream and co-expressed genes were related to plant hormones. The unique outcomes of our analyses indicate Potri.012G130000 is a promising candidate gene and

(GO:0044550), secondary metabolic process (GO:0019748), flavonoid biosynthetic process (GO:0009813), and flavonoid metabolic process (GO:0009812) (Fig. 3a). Notably, some of these downstream genes, such as Potri.007G019300 (closest Arabidopsis sequence homolog: AT5G66390), Potri.018G104800 (closest



**Fig. 3** Downstream and upstream expression quantitative trait loci (eQTL) analysis for single nucleotide polymorphism (SNP) Chr12:14 692 802 in *Populus trichocarpa*. (a) Gene Ontology (GO) enrichment analysis for the downstream genes of SNP Chr12:14 692 802 in xylem tissue. (b) Manhattan plot of eQTL analysis for the annotated gene Potri.012G130000 in xylem tissue. (c) GO enrichment analysis for target genes of upstream expression quantitative trait nucleotides (eQTNs) regulating gene Potri.012G130000 in xylem tissue. (d) Tree plot detailing GO enrichment analysis for genes co-expressed with Potri.012G130000 in xylem tissue.

that further experimental validation is merited to confirm its specific functions in lignocellulose biosynthesis.

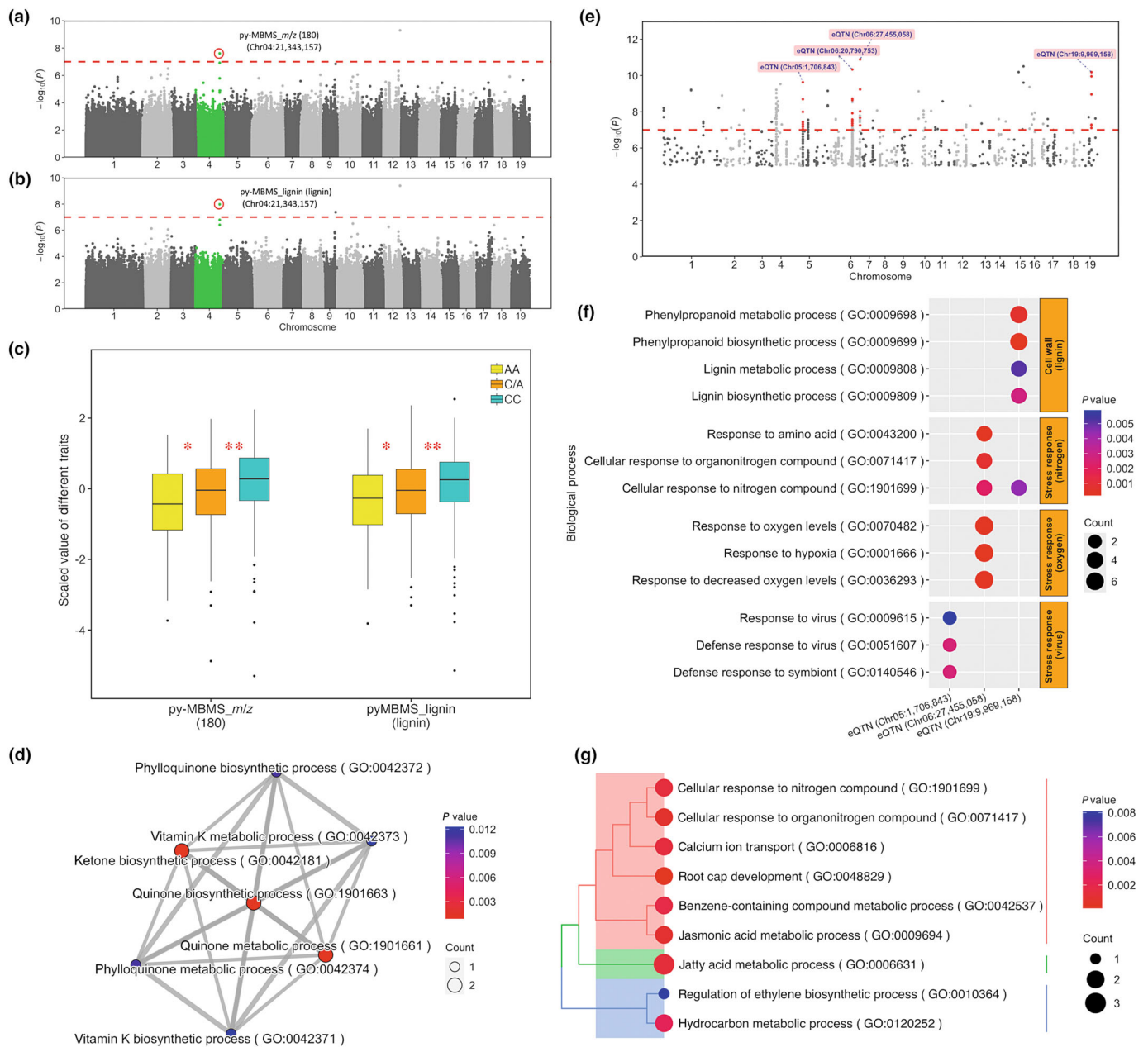
### Candidate gene associations with py-MBMS<sub>m/z</sub> and py-MBMS<sub>lignin</sub>

In addition to the SNP Chr12:14 692 802, we identified another significant SNP, Chr04:21 343 157, associated with both py-MBMS<sub>m/z</sub> (*m/z*: 180 (S, G)) and py-MBMS<sub>lignin</sub> (lignin content) (Table S2). This SNP displayed the lowest *P*-value association on Chromosome 4 for both traits (Fig. 4a,b). To further elucidate the relationship between SNP genotypes and the respective traits, we generated boxplots showcasing normalized measurements for both traits (Fig. 4c). The boxplots revealed that py-MBMS<sub>m/z</sub> (*m/z*: 180 (S, G)) and py-MBMS<sub>lignin</sub> (lignin content) consistently displayed the lowest values for the homozygous SNP genotype AA, the highest for genotype CC, and intermediate values for the heterozygous genotype C/A. This trend supports previous findings that associate py-MBMS<sub>m/z</sub> 180 (S, G) with lignin (Sykes *et al.*, 2008; Xiao *et al.*, 2014). Thus, the combined insights from the Manhattan plot and boxplots suggest a pivotal role for SNP Chr04:21 343 157 in lignin biosynthesis.

Next, we examined the downstream genes linked to SNP Chr04:21 343 157 in xylem tissue, drawing on *trans*-eQTL data. Our GO enrichment analysis revealed a significant enrichment of these genes in quinone and ketone biosynthesis and metabolic pathways (GO:1901663, GO:1901661, GO:0042374, GO:0042372, and GO:0042181) (Fig. 4d). Quinones are

integral to the lignin biosynthesis pathway (Boerjan *et al.*, 2003) and also act as defensive agents for herbivores and pathogens (Mittler, 2002; Constabel & Barbehenn, 2008). The identified enrichment suggests that the downstream genes of SNP Chr04:21 343 157 in xylem tissue may play a crucial role in modulating the biosynthesis of lignin, contributing both to structural integrity and defense mechanisms in xylem tissue.

SNP Chr04:21 343 157 is located in an intergenic region, flanked by genes Potri.004G202900 and Potri.004G203000. The closest Arabidopsis sequence homologs for both genes are identified as AT3G52740 (annotated as BIC1). To determine which of the two is most likely the causal locus, we utilized eQTL analysis with RNA expression data from xylem tissue to explore the upstream eQTNs associated with these genes. For Potri.004G202900, four significant upstream eQTNs ( $P < 10^{-7}$ ) were associated with its expression (Fig. 4e). Of these, only three eQTNs had a sufficient number of target genes to facilitate a GO enrichment analysis. The GO annotations for the target genes of these eQTNs, presented in Fig. 4(f), predominantly relate to cell wall biosynthesis and defense response pathways. This encompasses processes such as lignin biosynthesis (GO:0009698, GO:0009699, GO:0009808, and GO:0009809), stress responses to nitrogen (GO:0043200, GO:0071417, and GO:1901699), oxygen (GO:0070482, GO:0001666, and GO:0036293), and viruses (GO:0009615, GO:0051607, and GO:0140546). Additionally, genes co-expressed with Potri.004G202900 in xylem tissue are predominantly related to cellular responses to stimuli, including responses to nitrogen and the jasmonic acid metabolic



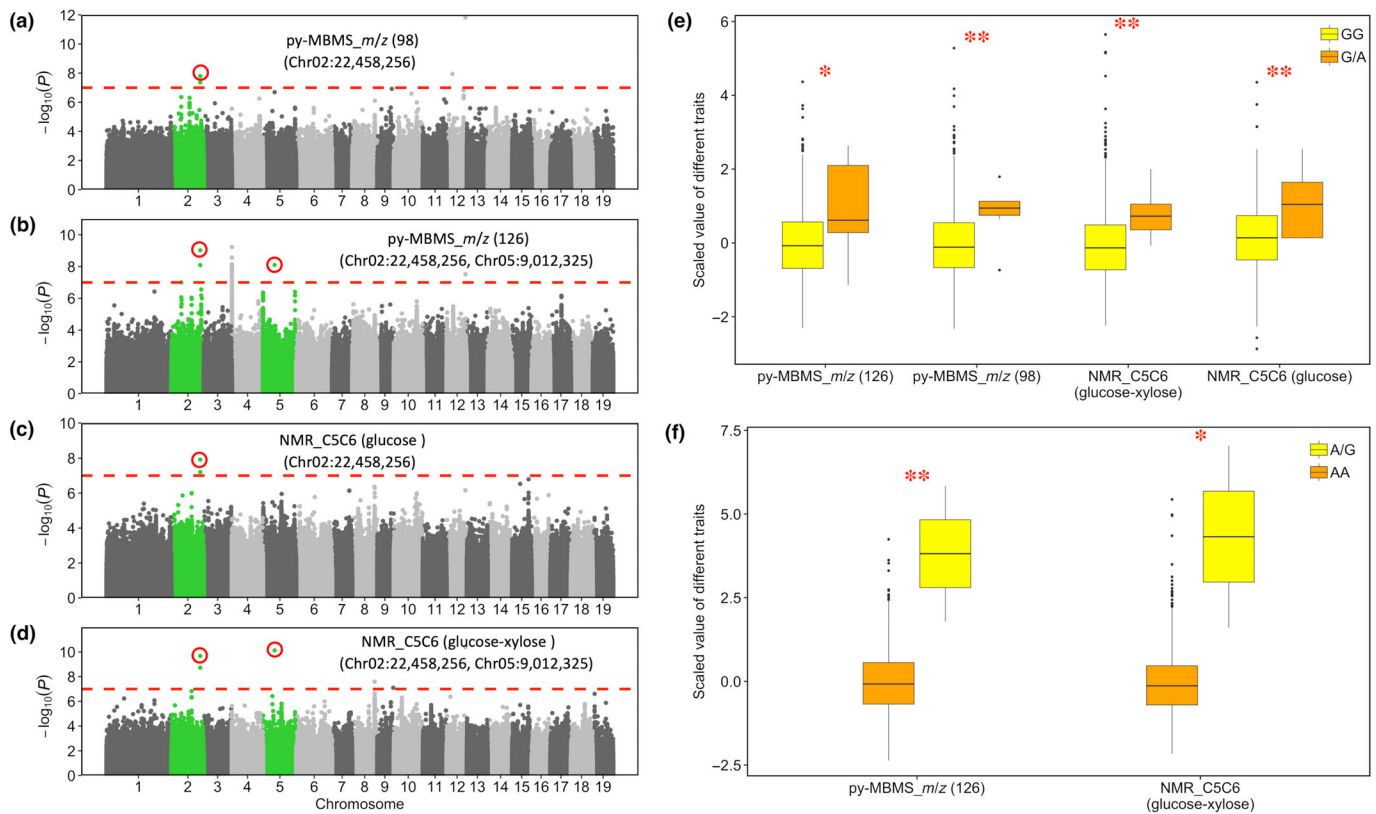
**Fig. 4** Comprehensive analysis of single nucleotide polymorphism (SNP) Chr04:21 343 157 associations and impacted genes in *Populus trichocarpa*. (a, b) Manhattan plots showcasing the genome-wide association studies (GWAS) signal at SNP Chr04:21 343 157, associated with py-MBMS\_m/z (*m/z*: 180) and py-MBMS\_lignin (lignin). (c) Boxplot illustrating traits significantly associated with SNP Chr04:21 343 157 across different traits. *Note*: Trait measurements are normalized. Horizontal lines within each box represent the median, the boxes represent the interquartile range (IQR), the whiskers extend to 1.5 times the IQR, and dots represent outliers. The characters \* and \*\* indicate *t*-test *P*-values < 0.05 and < 0.01, respectively. (d) Gene Ontology (GO) enrichment map for the downstream genes of SNP Chr04:21 343 157 in xylem tissue. (e) Manhattan plot of expression quantitative trait loci (eQTL) analysis for the annotated gene Potri.004G202900 in xylem tissue. (f) GO enrichment analysis for target genes of upstream expression quantitative trait nucleotides (eQTNs) regulating gene Potri.004G202900 in xylem tissue. (g) Tree plot detailing GO enrichment analysis for genes co-expressed with Potri.004G202900 in xylem tissue.

process (Fig. 4g). Conversely, for Potri.004G203000, neither its upstream nor co-expressed genes exhibited associations with cell wall or defense response functions. Based on these findings, we theorize that Potri.004G202900 (BIC1) is the most likely candidate gene associated with the SNP at Chr04:21 343 157 and is predicted to encode a Cryptochrome Function Inhibitor protein.

#### Candidate gene across py-MBMS\_m/z and NMR\_C5C6

We identified three significant SNPs associated with both py-MBMS\_m/z and NMR\_C5C6: SNP Chr02:22 458 248, SNP Chr02:22 458 256, and SNP Chr05:9012 325 (Table S2; Fig. 5a–d). Given that SNP Chr02:22 458 248 and SNP





**Fig. 5** Manhattan plots and boxplots for single nucleotide polymorphism (SNP) Chr02:22 458 256 and SNP Chr05:9012 325 in *Populus trichocarpa*. (a–d) Manhattan plot showcasing the genome-wide association studies (GWAS) signal at SNP Chr02:22 458 256 and SNP Chr05:9012 325 associated with py-MBMS\_m/z (*m/z*: 98, 126), NMR\_C5C6 (glucose, glucose-xylose). (e, f) Boxplot illustrating traits significantly associated with SNP Chr02:22 458 256 and SNP Chr05:9012 325 across different genotypes. Note: Trait measurements are scaled. Horizontal lines within each box represent the median, the boxes represent the interquartile range (IQR), the whiskers extend to 1.5 times the IQR, and dots represent outliers. The characters \* and \*\* indicate *t*-test *P*-values  $< 0.05$  and  $< 0.01$ , respectively.

Chr02:22 458 256 are proximal and within the same gene locus, we focused on SNP Chr02:22 458 256 and SNP Chr05:9012 325. These two SNPs exhibited the lowest *P*-value on Chromosomes 2 and 5, respectively (Table S6; Fig. 5a–d). Specifically, SNP Chr02:22 458 256 correlated with four traits, including py-MBMS\_m/z (*m/z*: 98, 126) and NMR\_C5C6 (glucose-xylose ratio, glucose). Boxplots revealed that the homozygous SNP genotype (GG) consistently had lower trait values compared with the heterozygous SNP genotype (G/A) across all four traits (Fig. 5e). Similarly, for SNP Chr05:9012 325, associated with py-MBMS\_m/z (*m/z*: 126 (C6)) and NMR\_C5C6 (glucose-xylose ratio), the homozygous genotype (AA) had lower values than the heterozygous genotype (A/G) (Fig. 5f).

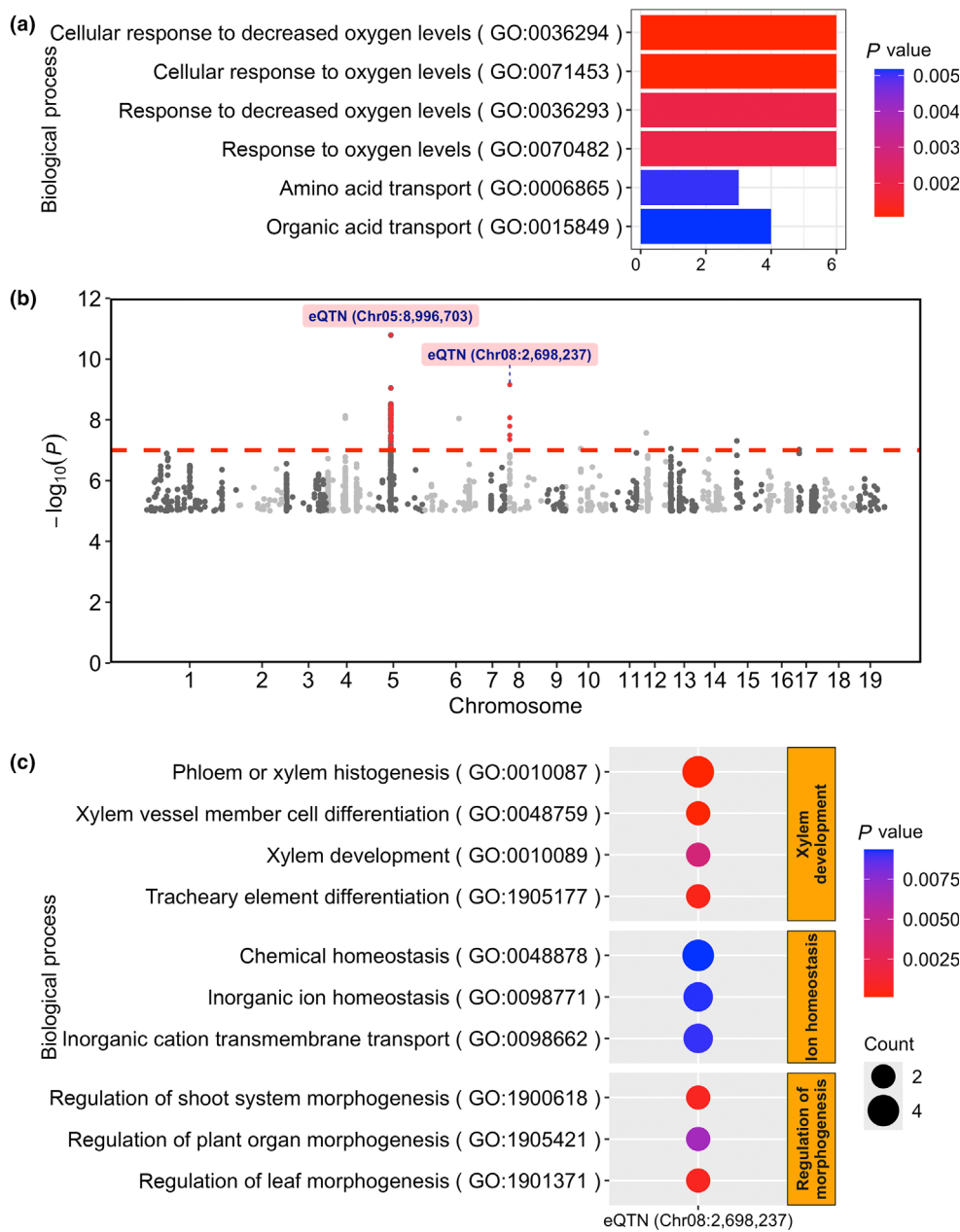
Interestingly, for SNP Chr02:22 458 256, the eQTL analysis did not identify any downstream genes in xylem tissues. This SNP is located between gene ID Potri.002G232100 (closest Arabidopsis sequence homolog: AT2G43080) and Potri.002G232200 (closest Arabidopsis sequence homolog: AT2G43070: SPPL3). No upstream eQTNs were detected for both genes. Consequently, we are unable to ascertain the exact annotation or function relevance of this SNP.

For SNP Chr05:9012 325, we identified downstream genes using the xylem eQTL data. The GO descriptions of these

genes predominantly pertained to stress responses to oxygen levels (GO:0071453, GO:0036294, GO:0036293, and GO:0070482) and amino acid transport (GO:0006865 and GO:0015849) (Fig. 6a).

SNP Chr05:9012 325 is located between two genes: Potri.005G116800 (with its closest Arabidopsis sequence homolog being AT2G18060: VND1) and Potri.005G116900 (closest Arabidopsis sequence homolog: AT3G51150). Notably, only gene Potri.005G116800 exhibited significant upstream eQTL nucleotides (eQTNs) ( $P < 10^{-7}$ ), including a *trans* upstream eQTN on Chromosome 8 and multiple *cis* regulators on Chromosome 5 (Fig. 6b). In xylem tissue, the primary functions of these upstream eQTN target genes encompass xylem development (GO:0010087, GO:0048759, GO:0010089, and GO:1905177), ion homeostasis (GO:0050801, GO:0098771, and GO:0006812), and morphogenesis regulation (GO:1900618, GO:1905421, and GO:1901371) (Fig. 6c).

Potri.005G116800 is annotated as a VND1 gene. Moreover, vascular-related NAC-domain (VND) genes have been identified as being pivotal in activating transcriptional regulators vital for secondary wall biosynthesis in plants (Zhou *et al.*, 2014; Zhang *et al.*, 2018a). Previous research confirms the important role of Potri.005G116800 (AT2G18060: VND1, PtrVND6-C2) as a



**Fig. 6** Downstream and upstream expression quantitative trait loci (eQTL) analysis for single nucleotide polymorphism (SNP) Chr05:9012 325 in xylem tissue. (a) Gene Ontology (GO) enrichment map for the downstream genes of SNP Chr05:9012 325 in xylem tissue. (b) Manhattan plot of eQTL analysis for the annotated gene Potri.005G116800 for SNP Chr05:9012 325 in xylem tissue. (c) GO enrichment analysis for target genes of upstream expression quantitative trait nucleotide (eQTN) regulating gene Potri.005G116800 in xylem tissue.

xylem-specific VND gene (Lin *et al.*, 2017; Takata *et al.*, 2019). Additionally, PtrVND6-C2 has been shown to not only activate its own expression but also that of other PtrVND genes (Lin *et al.*, 2017). This finding aligns with the *cis* regulators identified in our eQTL analysis for gene Potri.005G116800 (Fig. 6b). In summary, Potri.005G116800 (AT2G18060: VND1) emerges as the most probable candidate gene associated with SNP Chr05:9012 325.

#### Candidate gene across py-MBMS<sub>m/z</sub> and HSQC NMR

In GWAS analyses using both py-MBMS<sub>m/z</sub> and HSQC NMR lignin linkages, we identified five genomic regions that showed significant associations across both techniques (Table S6; Fig. 1). These regions, which we refer to as ‘co-occurring

intervals’, encompassed genes Potri.002G216300, Potri.004G000200 (most closely related *Arabidopsis* sequence homolog: AT2G01050), Potri.010G070800 (most closely related *Arabidopsis* sequence homolog: AT3G23280: XBAT35), Potri.T006300 (AT5G28780), and Potri.T006400 (AT2G01050). Although these intervals did not always involve identical SNPs, they were located within a 10-kb segment, allowing us to identify a consistent pattern of genetic association across the alternate analytical techniques.

Examining these five genes, we found that Potri.002G216300 was significantly associated with three traits, including py-MBMS<sub>m/z</sub> (*m/z*: 60 (C5, C6), 73 (C5, C6)) and HSQC NMR (resinol linkage) (Table S2). Similarly, Potri.004G000200 (AT2G01050) displayed significant associations with py-MBMS<sub>m/z</sub> (*m/z*: 126 (C6)) and HSQC NMR (p-

hydroxyphenyl) (Table S6). However, eQTL analysis in xylem tissue did not identify any downstream or upstream genes for either gene, and no co-expressed genes were found. Despite their co-occurrence in GWAS analysis using two distinct phenotyping methods, the importance of these genes remains unclear at this stage.

The remaining three co-occurring genes, Potri.010G070800 (AT3G23280: XBAT35), Potri.T006300 (AT5G28780), and Potri.T006400 (AT2G01050), each exhibited significant associations with two traits: py-MBMS\_*m/z* (*m/z*: 60 (C5, C6)) and HSQC NMR (p-hydroxyphenyl) (Table S6). For gene Potri.010G070800 (AT3G23280: XBAT35), we analyzed downstream and co-expressed genes in xylem tissue. The GO functions of both co-expressed and downstream genes were mainly involved in cell wall biogenesis and lignin metabolic processes (GO:0042546, GO:0009832, GO:0009834, GO:0010087, GO:0009808, GO:0010088, GO:0009698, and GO:0009699) (Fig. S5). These findings underscore the pivotal role of Potri.010G070800 in GWAS analysis and its implications in stress response and cell wall formation, corroborating similar observations reported in recent research (Yao *et al.*, 2023).

As for Potri.T006300 (AT5G28780) and Potri.T006400 (AT2G01050), no downstream or upstream genes were discovered through eQTL analysis. However, Potri.T006300 (AT5G28780) was co-expressed with genes primarily associated with the stress response to hypoxia and salicylic acid in the xylem tissue (GO:0001666, GO:0070482, and GO:0009751) (Fig. S5). These observations underline the potential role of Potri.T006300 in lignin biosynthesis, possibly as part of a defense response.

In total, 598 SNPs (Table S6) were significantly associated with various phenotypic traits measured in our study. After applying the overlapping associations with different phenotyping methods standard, as well as the eQTL and co-expression criteria, only Potri.005G116800, Potri.012G130000, and Potri.004G202900 satisfied all three integrated criteria. These candidate genes demonstrated strong potential involvement in lignocellulose biosynthesis, warranting further experimental validation to confirm their specific roles.

## Discussion

In this study, we addressed a major limitation commonly associated with GWAS: the lack of biological context in single loci associations that link nucleotide variation to traits. To overcome this limitation, we employed an integrative multiple-measurement approach. This methodology enabled us to gain biological insights through the transcriptional and co-expression networks associated with SNPs identified in our GWAS. Specifically, we assessed four distinct sets of lignocellulose traits using three methods: py-MBMS (*m/z* & lignin content), HSQC NMR (lignin linkages), and NMR (cell wall carbohydrate). Our GWAS analyses identified a multitude of genetic variants linked to lignin and carbohydrate traits. To refine our focus, we filtered these variants to retain only those linked with two or more sets of lignocellulose traits. Furthering our investigation, eQTL analysis was

employed to clarify putative network and regulatory targets, associated upstream genes, and co-expressed genes of these variants, thus enhancing the biological interpretation of the GWAS results. Our findings suggest that eQTL mapping offers a high-throughput approach linking genetic variants to gene expression in lignocellulose development. The integration of GWAS and eQTL mapping in our study sheds new light on the genetic regulation of lignocellulose and emphasizes the crucial role of several SNP-associated genes distributed across different chromosomes. By annotating these SNPs, we highlighted three key genes (Potri.005G116800, Potri.012G130000, and Potri.004G202900), two that are novel candidate genes potentially involved in the biosynthetic network and/or regulation of lignocellulose development.

Our study found that py-MBMS\_lignin (lignin content) was significantly associated with two SNPs, both of which also exhibited strong correlations with traits measured by other methods in GWAS analyses. Specifically, SNP Chr12:14,692,802 was associated with nine py-MBMS\_*m/z* traits (*m/z*: 57 (C5, C6), 60 (C5, C6), 73 (C5, C6), 98 (C6), 126 (C6), 144 (C6), 137 (G), 154 (S), 180 (S, G)) and one NMR\_C5C6 trait (glucose–xylose ratio). Another SNP, Chr04:21,343,157, was associated with one py-MBMS\_*m/z* (*m/z*: 180 (S, G)) trait. However, no SNP was significantly associated with the S/G ratio in py-MBMS\_lignin in the GWAS analyses. S/G ratio determination in py-MBMS\_lignin creates relative S/G rankings (high/low) but not absolute S/G ratio values (Sykes *et al.*, 2008). It is likely that by computing S/G ratio from the py-MBMS\_*m/z* value, we either overestimated or underestimated the natural variation in the S/G ratio in *P. trichocarpa* and thus found no significant associations.

Trait variation assessed by py-MBMS\_*m/z* encompasses not only the variation in lignin but also carbohydrates. The carbohydrate phenotypes measured via py-MBMS\_*m/z* were the only traits exhibiting co-occurring SNPs with all three other datasets in the GWAS analysis. These observations highlight the comprehensive nature of the py-MBMS\_*m/z* method in capturing variations in cellulosic traits. However, the primary GWAS limitations of py-MBMS\_*m/z* involve challenges in interpreting different *m/z* values. Many *m/z* originate from multiple chemical species and plant cell wall components, leading to a substantial number of significant associations with over 400 *m/z* values. By contrast, phenotypic variation measured by HSQC NMR identified the highest number of significantly associated genes (490) despite using the fewest number of samples and phenotypes. This might be indicative of higher rates of false positives given the smaller population size. Bryant *et al.* (2023) have suggested that although HSQC NMR is more labor- and time-intensive, it yields considerably more information on lignin composition and structure.

Among these genes reported in our study, only Potri.005G116800 (AT2G18060: VND1) has previously been established as a crucial regulator of lignocellulose biosynthesis in *Populus* (Lin *et al.*, 2017; Takata *et al.*, 2019; Akiyoshi *et al.*, 2021). Our GWAS results support the importance of this gene in cell wall formation. Subsequent eQTL analysis suggested that this gene may also regulate genes involved in defense against

various stimuli. For example, we uncovered defense responses (e.g. response to oxygen level) within the GO functions of downstream genes in xylem tissue for Potri.005G116800. Additionally, the upstream GO terms appear to be involved in both xylem development and stress responses. Consistent with our findings, prior co-expression studies have indicated that the expression and interaction patterns of VND family genes could differ under stress conditions (Taylor-Teeples *et al.*, 2015; Ohtani & Demura, 2019). Molecular genetic investigations have further revealed that VND family protein activity can be dynamically regulated in response to light (Tan *et al.*, 2018) and cellular thiol (Kawabe *et al.*, 2018; Ohtani *et al.*, 2018). Numerous studies have established the crucial role of cell walls in plant resistance to diverse stressors (Hamann, 2012; Miedes *et al.*, 2014; Houston *et al.*, 2016). Hamann (2012) suggested that plant cell walls aid in perceiving and transducing signals to activate the defense response, primarily through alterations in composition and structure. Bellincampi *et al.* (2014) reported that the ability of plants to counteract pathogen-produced cell wall-degrading enzymes is reinforced via lignin biosynthesis. Collectively, these findings illustrate the intricate connections between cell wall properties and defense mechanisms in trees. Our study, when combined with prior research, supports the role of the Potri.005G116800 (AT2G18060: VND1) in regulating cell wall biosynthesis and promoting the expression of defense response-related genes.

Among the remaining two genes with significant roles in lignocellulose formation – Potri.012G130000 (AT4G22820: AtSAP9) and Potri.004G202900 (AT3G52740: BIC1) – each has identified upstream and downstream networks connected with lignocellulose biosynthesis, validating their contributions to this process. Potri.012G130000 predominantly downregulated genes integral to lignin biosynthesis. Its upstream and co-expressed genes predominantly function in the defense response in xylem tissue. For instance, the upstream genes of Potri.012G130000 are linked to hormone biosynthesis, supporting prior research reporting on the role of hormone regulation in wood formation and defense response in *Populus* (Immanen *et al.*, 2016; Sundell *et al.*, 2017). The downstream genes in xylem for Potri.004G202900 are mainly involved in quinones and ketones biosynthesis, which have been reported to be integral in the lignin biosynthesis pathway (Boerjan *et al.*, 2003), and act as defensive agents for herbivores and pathogens (Mittler, 2002; Constabel & Barbehenn, 2008). The upstream genes of Potri.004G202900 are enriched for lignin biosynthesis and defense response pathways. Intriguingly, its co-expressed genes in xylem chiefly correspond to genes that respond to multiple stimuli; for example, AT3G52740 (BIC1: blue-light inhibitor of cryptochromes 1) and Potri.004G202900, a member of the BRASSINAZOLE-RESISTANT 1 (BZR1) transcription factor family. BIC1 has been documented to augment brassinosteroid signaling, fostering cell elongation and bolstering plant immune responses (Tang *et al.*, 2016; Yang *et al.*, 2021). In short, the integration of our GWAS analyses, eQTL examination of upstream and downstream networks, and co-expression analyses collectively indicate the potential significance of Potri.012G130000 and Potri.004G202900 in lignocellulose formation, particularly

concerning lignocellulose biosynthesis with a likely role in defense responses.

In conclusion, our integrated GWAS and eQTL analyses, supported by co-expression analysis, led to the identification of three key candidate genes potentially involved in lignocellulose biosynthesis in *Populus*. These findings expand our understanding of the complex molecular network underlying cell wall biosynthesis and defense responses in poplar trees. The two newly identified genes (Potri.012G130000 and Potri.004G202900) are promising candidates for further investigation and experimental validation to confirm their roles. Our study demonstrates the effectiveness of combining these different analytical approaches in investigating complex biological processes, particularly when comprehensive gene network information is unavailable. This integrative methodology provides potential targets for further investigation and manipulation to improve wood properties and stress resistance in *Populus* and other tree species.

## Acknowledgements

This material is based upon work supported by the Center for Bioenergy Innovation (CBI), US Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number ERKP886. This research used resources of the Compute and Data Environment for Science (CADES) and the Oak Ridge Leadership Computing Facility (OLCF). Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the Office of Science of the U.S. Department of Energy under contract no.: DE-AC05-00OR22725. The work (proposal: 10.46936/10.25585/60001221) conducted by the US Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under contract no.: DE-AC02-05CH11231.

## Competing interests

None declared.

## Author contributions

MS, J-GC, WM and GAT conceived and designed the project. AEH, RMH, NB, SSJ and ATR generated and analyzed the experimental data. MS, TBY and CJ conducted the bioinformatic analysis. MS, J-GC and GAT wrote the manuscript with contributions from AEH, RMH and WM. All authors critically reviewed and approved the manuscript.

## ORCID

Nathan Bryant  <https://orcid.org/0000-0003-0004-9448>  
 Jin-Gui Chen  <https://orcid.org/0000-0002-1752-4201>  
 Renee M. Happs  <https://orcid.org/0000-0001-7139-0083>  
 Anne E. Harman-Ware  <https://orcid.org/0000-0002-7927-9424>  
 Sara S. Jawdy  <https://orcid.org/0000-0002-8123-5439>

Wellington Muchero  <https://orcid.org/0000-0002-0200-9856>

Arthur J. Ragauskas  <https://orcid.org/0000-0002-3536-554X>

Mengjun Shu  <https://orcid.org/0000-0002-6323-2664>

Gerald A. Tuskan  <https://orcid.org/0000-0003-0106-1289>

## Data availability

Data are available at doi: [10.25983/2352478](https://doi.org/10.25983/2352478).

## References

- Akiyoshi N, Ihara A, Matsumoto T, Takebayashi A, Hiroyama R, Kikuchi J, Demura T, Ohtani M. 2021. Functional analysis of poplar *sombrero*-type NAC transcription factors yields a strategy to modify woody cell wall properties. *Plant and Cell Physiology* 62: 1963–1974.
- Alseekh S, Kostova D, Bulut M, Fernie AR. 2021. Genome-wide association studies: assessing trait characteristics in model and crop plants. *Cellular and Molecular Life Sciences* 78: 5743–5754.
- Bellincampi D, Cervone F, Lionetti V. 2014. Plant cell wall dynamics and wall-related susceptibility in plant–pathogen interactions. *Frontiers in Plant Science* 5: 228.
- Bhat JA, Yu D. 2021. High-throughput NGS-based genotyping and phenotyping: role in genomics-assisted breeding for soybean improvement. *Legume Science* 3: e81.
- Boerjan W, Ralph J, Baucher M. 2003. Lignin biosynthesis. *Annual Review of Plant Biology* 54: 519–546.
- Bryant N, Zhang J, Feng K, Shu M, Ployet R, Chen J-G, Muchero W, Yoo CG, Tschaplinski TJ, Pu Y *et al.* 2023. Novel candidate genes for lignin structure identified through genome-wide association study of naturally varying *Populus trichocarpa*. *Frontiers in Plant Science* 14: 3113.
- Buttò V, Deslauriers A, Rossi S, Rozenberg P, Shishov V, Morin H. 2020. The role of plant hormones in tree-ring formation. *Trees* 34: 315–335.
- Carlson M. 2021. *org.At.tair.db: genome wide annotation for Arabidopsis*. R package v.3.8.2. [WWW document] URL <https://bioconductor.org/packages/release/data/annotation/html/org.At.tair.db.html> [accessed 30 January 2024].
- Chhetri HB, Macaya-Sanz D, Kainer D, Biswal AK, Evans LM, Chen J-G, Collins C, Hunt K, Mohanty SS, Rosenstiel T *et al.* 2019. Multitrait genome-wide association analysis of *Populus trichocarpa* identifies key polymorphisms controlling morphological and physiological traits. *New Phytologist* 223: 293–309.
- Constabel CP, Barbehenn R. 2008. Defensive roles of polyphenol oxidase in plants. In: Schaller A, ed. *Induced plant resistance to herbivory*. Dordrecht, the Netherlands: Springer, 253–270.
- D'Agostino N, Tripodi P. 2017. NGS-based genotyping, high-throughput phenotyping and genome-wide association studies laid the foundations for next-generation breeding in horticultural crops. *Diversity* 9: 38.
- Davis MF, Tuskan GA, Payne P, Tschaplinski TJ, Meilan R. 2006. Assessment of *Populus* wood chemistry following the introduction of a Bt toxin gene. *Tree Physiology* 26: 557–564.
- Demirjian C, Vaillau F, Berthomé R, Roux F. 2023. Genome-wide association studies in plant pathosystems: success or failure? *Trends in Plant Science* 28: 471–485.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21.
- Eckert C, Sharmin S, Kogel A, Yu D, Kins L, Strijkstra G-J, Polle A. 2019. What makes the wood? Exploring the molecular mechanisms of xylem acclimation in hardwoods to an ever-changing environment. *Forests* 10: 358.
- Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G *et al.* 2014. Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nature Genetics* 46: 1089–1096.
- Feder ME, Mitchell-Olds T. 2003. Evolutionary and ecological functional genomics. *Nature Reviews Genetics* 4: 649–655.
- Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyster AE, Denny JC, Nicolae DL, Cox NJ *et al.* 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* 47: 1091–1098.
- Groover AT, Nieminen K, Helariutta Y, Mansfield SD. 2010. Wood formation in *Populus*. In: Jansson S, Bhalarao R, Groover A, eds. *Plant genetics and genomics: crops and models. Genetics and genomics of Populus*. New York, NY, USA: Springer, 201–224.
- Guerra FP, Suren H, Holliday J, Richards JH, Fiehn O, Famula R, Stanton BJ, Shuren R, Sykes R, Davis MF *et al.* 2019. Exome resequencing and GWAS for growth, ecophysiology, and chemical and metabolomic composition of wood of *Populus trichocarpa*. *BMC Genomics* 20: 875.
- Guerra FP, Wegrzyn JL, Sykes R, Davis MF, Stanton BJ, Neale DB. 2013. Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytologist* 197: 162–176.
- Hamann T. 2012. Plant cell wall integrity maintenance as an essential component of biotic stress response mechanisms. *Frontiers in Plant Science* 3: 77.
- Hamilton JP, Robin BC. 2012. Advances in plant genome sequencing. *The Plant Journal* 70: 177–190.
- Happs RM, Bartling AW, Doepfke C, Harman-Ware AE, Clark R, Webb EG, Biddy MJ, Chen J-G, Tuskan GA, Davis MF *et al.* 2021. Economic impact of yield and composition variation in bioenergy crops: *Populus trichocarpa*. *Biofuels, Bioproducts and Biorefining* 15: 176–188.
- Harman-Ware AE, Happs RM, Macaya-Sanz D, Doepfke C, Muchero W, DiFazio SP. 2022. Abundance of major cell wall components in natural variants and pedigrees of *Populus trichocarpa*. *Frontiers in Plant Science* 13: 757810.
- Harman-Ware AE, Macaya-Sanz D, Abeyratne CR, Doepfke C, Haiby K, Tuskan GA, Stanton B, DiFazio SP, Davis MF. 2021. Accurate determination of genotypic variance of cell wall characteristics of a *Populus trichocarpa* pedigree using high-throughput pyrolysis-molecular beam mass spectrometry. *Biotechnology for Biofuels* 14: 59.
- Houston K, Tucker MR, Chowdhury J, Shirley N, Little A. 2016. The plant cell wall: a complex and dynamic structure as revealed by the responses of genes under stress conditions. *Frontiers in Plant Science* 7: 984.
- Immanen J, Nieminen K, Smolander O-P, Kojima M, Alonso Serra J, Koskinen P, Zhang J, Elo A, Mähönen AP, Street N *et al.* 2016. Cytokinin and auxin display distinct but interconnected distribution and signaling profiles to stimulate cambial activity. *Current Biology* 26: 1990–1997.
- Ingvarsson PK, Street NR. 2011. Association genetics of complex traits in plants. *New Phytologist* 189: 909–922.
- Kawabe H, Ohtani M, Kurata T, Sakamoto T, Demura T. 2018. Protein S-nitrosylation regulates xylem vessel cell differentiation in *Arabidopsis*. *Plant & Cell Physiology* 59: 17–29.
- Korte A, Farlow A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9: 29.
- Kumar M, Campbell L, Turner S. 2016. Secondary cell walls: biosynthesis and manipulation. *Journal of Experimental Botany* 67: 515–531.
- Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 559.
- Li Y, Jin F, Chao Q, Wang B-C. 2017. Proteomics analysis reveals the molecular mechanism underlying the transition from primary to secondary growth of poplar. *Journal of Plant Physiology* 213: 1–15.
- Li Z, Wang P, You C, Yu J, Zhang X, Yan F, Ye Z, Shen C, Li B, Guo K *et al.* 2020. Combined GWAS and eQTL analysis uncovers a genetic regulatory network orchestrating the initiation of secondary cell wall development in cotton. *New Phytologist* 226: 1738–1752.
- Liao Y, Smyth GK, Shi W. 2014. FEATURECOUNTS: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30: 923–930.
- Lin Y-CJ, Chen H, Li Q, Li W, Wang JP, Shi R, Tunlaya-Anukit S, Shuai P, Wang Z, Ma H *et al.* 2017. Reciprocal cross-regulation of VND and SND multigene TF families for wood formation in *Populus trichocarpa*. *Proceedings of the National Academy of Sciences, USA* 114: E9722–E9729.

- Mähler N, Schiffthaler B, Robinson KM, Terebieniec BK, Vučak M, Mannapperuma C, Bailey MES, Jansson S, Hvidsten TR, Street NR. 2020. Leaf shape in *Populus tremula* is a complex, omnigenic trait. *Ecology and Evolution* 10: 11922–11940.
- Marriott PE, Gómez LD, McQueen-Mason SJ. 2016. Unlocking the potential of lignocellulosic biomass through plant science. *New Phytologist* 209: 1366–1381.
- Meents MJ, Watanabe Y, Samuels AL. 2018. The cell biology of secondary cell wall biosynthesis. *Annals of Botany* 121: 1107–1125.
- Miedes E, Vanholme R, Boerjan W, Molina A. 2014. The role of the secondary cell wall in plant resistance to pathogens. *Frontiers in Plant Science* 5: 358.
- Mittler R. 2002. Oxidative stress, antioxidants and stress tolerance. *Trends in Plant Science* 7: 405–410.
- Muchero W, Guo J, DiFazio SP, Chen J-G, Ranjan P, Slavov GT, Gunter LE, Jawdy S, Bryan AC, Sykes R *et al.* 2015. High-resolution genetic mapping of allelic variants associated with cell wall chemistry in *Populus*. *BMC Genomics* 16: 24.
- Ohtani M, Demura T. 2019. The quest for transcriptional hubs of lignin biosynthesis: beyond the NAC-MYB-gene regulatory network model. *Current Opinion in Biotechnology* 56: 82–87.
- Ohtani M, Kawabe H, Demura T. 2018. Evidence that thiol-based redox state is critical for xylem vessel cell differentiation. *Plant Signaling & Behavior* 13: e1428512.
- Porter HF, O'Reilly PF. 2017. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Scientific Reports* 7: 38837.
- Porth I, El-Kassaby YA. 2015. Using *Populus* as a lignocellulosic feedstock for bioethanol. *Biotechnology Journal* 10: 510–524.
- Ren M, Zhang Y, Wang R, Liu Y, Li M, Wang X, Chen X, Luan X, Zhang H, Wei H *et al.* 2022. *PttHAT22*, as a higher hierarchy regulator, coordinately regulates secondary cell wall component biosynthesis in *Populus trichocarpa*. *Plant Science* 316: 111170.
- Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. 2015. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* 16: 85–97.
- Robinson MD, McCarthy DJ, Smyth GK. 2010. EDGER: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26: 139–140.
- Rubin EM. 2008. Genomics of cellulosic biofuels. *Nature* 454: 841–845.
- Sannigrahi P, Ragauskas AJ, Tuskan GA. 2010. Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels, Bioproducts and Biorefining* 4: 209–226.
- Studer MH, DeMartini JD, Davis MF, Sykes RW, Davison B, Keller M, Tuskan GA, Wyman CE. 2011. Lignin content in natural *Populus* variants affects sugar release. *Proceedings of the National Academy of Sciences, USA* 108: 6300–6305.
- Su W-L, Liu N, Mei L, Luo J, Zhu Y-J, Liang Z. 2019. Global transcriptomic profile analysis of genes involved in lignin biosynthesis and accumulation induced by boron deficiency in poplar roots. *Biomolecules* 9: 156.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al.* 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, USA* 102: 15545–15550.
- Sundell D, Street NR, Kumar M, Mellerowicz EJ, Kucukoglu M, Johnsson C, Kumar V, Mannapperuma C, Delhomme N, Nilsson O *et al.* 2017. ASPWOOD: high-spatial-resolution transcriptome profiles reveal uncharacterized modularity of wood formation in *Populus tremula*. *Plant Cell* 29: 1585–1604.
- Sykes R, Krodzkycki B, Tuskan G, Foutz K, Davis M. 2008. Within tree variability of lignin composition in *Populus*. *Wood Science and Technology* 42: 649–661.
- Syvänen A-C. 2005. Toward genome-wide SNP genotyping. *Nature Genetics* 37: S5–S10.
- Takata N, Awano T, Nakata MT, Sano Y, Sakamoto S, Mitsuda N, Taniguchi T. 2019. *Populus* NST/SND orthologs are key regulators of secondary cell wall formation in wood fibers, phloem fibers and xylem ray parenchyma cells. *Tree Physiology* 39: 514–525.
- Tan TT, Endo H, Sano R, Kurata T, Yamaguchi M, Ohtani M, Demura T. 2018. Transcription factors VND1–VND3 contribute to cotyledon xylem vessel formation. *Plant Physiology* 176: 773–789.
- Tang J, Han Z, Chai J. 2016. Q&A: what are brassinosteroids and how do they act in plants? *BMC Biology* 14: 113.
- Taylor G. 2002. *Populus*: Arabidopsis for forestry. Do we need a model tree? *Annals of Botany* 90: 681–689.
- Taylor-Teeples M, Lin L, de Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Velling MT, Lamothe R *et al.* 2015. An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature* 517: 571–575.
- Tibbs Cortes L, Zhang Z, Yu J. 2021. Status and prospects of genome-wide association studies in plants. *The Plant Genome* 14: e20077.
- Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596–1604.
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics* 101: 5–22.
- Wegrzyn JL, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R, Davis MF, Tsai C-J, Neale DB. 2010. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. *New Phytologist* 188: 515–532.
- Wei T, Simko V. 2017. R package ‘CORRPLLOT’: visualization of a correlation matrix (v.0.84). [WWW document] URL <https://github.com/taiyun/corrplot> [accessed 30 January 2024].
- Weigel D, Nordborg M. 2015. Population genomics for understanding adaptation in wild plant species. *Annual Review of Genetics* 49: 315–338.
- Welker CM, Balasubramanian VK, Petti C, Rai KM, DeBolt S, Mendu V. 2015. Engineering plant biomass lignin content and composition for biofuels and bioproducts. *Energies* 8: 7654–7676.
- Wickham H. 2011. ggplot2. *WIREs Computational Statistics* 3: 180–185.
- Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L *et al.* 2021. CLUSTERPROFILER 4.0: a universal enrichment tool for interpreting omics data. *The Innovations* 2: 100141.
- Wullschlegel SD, Jansson S, Taylor G. 2002. Genomics and forest biology: *Populus* emerges as the perennial favorite. *Plant Cell* 14: 2651–2655.
- Xiao L, Wei H, Himmel ME, Jameel H, Kelley SS. 2014. NIR and Py-mbms coupled with multivariate data analysis as a high-throughput biomass characterization technique: a review. *Frontiers in Plant Science* 5: 388.
- Xu Z, Wu C, Wei P, Pan W. 2017. A powerful framework for integrating eQTL and GWAS summary data. *Genetics* 207: 893–902.
- Yang Z, Yan B, Dong H, He G, Zhou Y, Sun J. 2021. BIC1 acts as a transcriptional coactivator to promote brassinosteroid signaling and plant growth. *EMBO Journal* 40: e104615.
- Yao T, Zhang J, Yates TB, Shrestha HK, Engle NL, Ployet R, John C, Feng K, Bewg WP, Chen MSS *et al.* 2023. Expression quantitative trait loci mapping identified *PttXB38* as a key hub gene in adventitious root development in *Populus*. *New Phytologist* 239: 2248–2264.
- Yates TB, Feng K, Zhang J, Singan V, Jawdy SS, Ranjan P, Abraham PE, Barry K, Lipzen A, Pan C *et al.* 2021. The ancient salicoid genome duplication event: a platform for reconstruction of *de novo* gene evolution in *Populus trichocarpa*. *Genome Biology and Evolution* 13: evab198.
- Yu G. 2022. ENRICHPLOT: visualization of functional enrichment result. [WWW document] URL <https://bioconductor.org/packages/release/bioc/html/enrichplot.html> [accessed 30 January 2024].
- Zhang J, Tuskan GA, Tschaplinski TJ, Muchero W, Chen J-G. 2020. Transcriptional and post-transcriptional regulation of lignin biosynthesis pathway genes in *Populus*. *Frontiers in Plant Science* 11: 652.
- Zhang J, Xie M, Tuskan GA, Muchero W, Chen J-G. 2018a. Recent advances in the transcriptional regulation of secondary cell wall biosynthesis in the woody plants. *Frontiers in Plant Science* 9: 1535.
- Zhang J, Yang Y, Zheng K, Xie M, Feng K, Jawdy SS, Gunter LE, Ranjan P, Singan VR, Engle N *et al.* 2018b. Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT2 in caffeoylquinic acid biosynthesis and its regulation by

- defense-responsive transcription factors in *Populus*. *New Phytologist* 220: 502–516.
- Zhou J, Zhong R, Ye Z-H. 2014. Arabidopsis NAC domain proteins, VND1 to VND5, are transcriptional regulators of secondary wall biosynthesis in vessels. *PLoS ONE* 9: e105726.
- Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44: 821–824.
- Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM *et al.* 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* 48: 481–487.

## Supporting Information

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

**Fig. S1** Distribution of 47 phenotypic measurements across 834 *Populus trichocarpa* genotypes.

**Fig. S2** Pearson correlation matrix of 47 phenotypes.

**Fig. S3** Q-Q plot for GWAS results of 47 phenotypes.

**Fig. S4** Cluster dendrogram of co-expressed gene modules in xylem tissues of *Populus trichocarpa*.

**Fig. S5** Co-expressed and downstream genes for Potri.010G070800 in xylem tissue.

**Methods S1** SNP calling protocol for whole-genome DNA short-read.

**Table S1** Four sets of phenotypes (47 in total) measured in HSQC NMR, py-MBMS, and  $^1\text{H}$  NMR techniques in *Populus trichocarpa*.

**Table S2** Phenotypic measurements (47 in total) for 834 *Populus trichocarpa* genotypes utilizing HSQC NMR, py-MBMS, and  $^1\text{H}$  NMR techniques.

**Table S3** Genotype IDs, trait measurements, and RNA-seq information for the 1323 genotypes in the GWAS population in *Populus trichocarpa*.

**Table S4** NCBI SRA accession numbers and genotype IDs for RNA-Seq data in *Populus trichocarpa*.

**Table S5** GWAS results for all 47 phenotypes including all SNPs with  $P < 10^{-7}$  in *Populus trichocarpa*.

**Table S6** Thirteen genes associated with at least two different sets of phenotypes measurements via GWAS analysis in *Populus trichocarpa*.

Please note: Wiley is 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.