Lawrence Berkeley National Laboratory

Title

Pleiotropic and Epistatic Network-Based Discovery: Integrated Networks for Target Gene Discovery

Permalink

https://escholarship.org/uc/item/0fn4k8dw

Authors

Weighill, Deborah
Jones, Piet
Shah, Manesh
et al.

Publication Date

2018

DOI

10.3389/fenrg.2018.00030

Peer reviewed

Pleiotropic and Epistatic Network-Based Discovery: Integrated Networks for Target Gene Discovery

Deborah Weighill[1,2], Piet Jones[1,2], Manesh Shah[2], Priya Ranjan[2,3], Wellington Muchero[2], Jeremy Schmutz[4,5], Avinash Sreedasyam[5], David Macaya-Sanz[6], Robert Sykes[7], Nan Zhao[3], Madhavi Z. Martin[2], Stephen DiFazio[6], Timothy J. Tschaplinski[2], Gerald Tuskan[2] and Daniel Jacobson[1,2*]

[1] The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN, United States, [2] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [3] Department of Plant Sciences, The University of Tennessee Institute of Agriculture, University of Tennessee, Knoxville, TN, United States, [4] Department of Energy, Joint Genome Institute, Walnut Creek, CA, United States, [5] HudsonAlpha Institute for Biotechnology, Huntsville, AL, United States, [6] Department of Biology, West Virginia University, Morgantown, WV, United States, [7] National Renewable Energy Laboratory, Golden, CO, United States

Biological organisms are complex systems that are composed of functional networks of interacting molecules and macro-molecules. Complex phenotypes are the result of orchestrated, hierarchical, heterogeneous collections of expressed genomic variants. However, the effects of these variants are the result of historic selective pressure and current environmental and epigenetic signals, and, as such, their co-occurrence can be seen as genome-wide correlations in a number of different manners. Biomass recalcitrance (i.e., the resistance of plants to degradation or deconstruction, which ultimately enables access to a plant's sugars) is a complex polygenic phenotype of high importance to biofuels initiatives. This study makes use of data derived from the re-sequenced genomes from over 800 different *Populus trichocarpa* genotypes in combination with metabolomic and pyMBMS data across this population, as well as co-expression and co-methylation networks in order to better understand the molecular interactions involved in recalcitrance, and identify target genes involved in lignin biosynthesis/degradation. A Lines Of Evidence (LOE) scoring system is developed to integrate the information in the different layers and quantify the number of lines of evidence linking genes to target functions. This new scoring system was applied to quantify the lines of evidence linking genes to lignin-related genes and phenotypes across the network layers, and allowed for the generation of new hypotheses surrounding potential new candidate genes involved in lignin biosynthesis in *P. trichocarpa*, including various AGAMOUS-LIKE genes. The resulting Genome Wide Association Study networks, integrated with Single Nucleotide Polymorphism (SNP) correlation, co-methylation, and co-expression networks through the LOE scores are proving to be a powerful approach to determine the pleiotropic and epistatic relationships underlying cellular functions and, as such, the molecular basis for complex phenotypes, such as recalcitrance.

Keywords: multi-omic data layering, LOE scores, lines of evidence, GWAS, SNP correlation, association networks

1. INTRODUCTION

Populus species are promising sources of cellulosic biomass for biofuels because of their fast growth rate, high cellulose content and moderate lignin content (Sannigrahi et al., 2010). Ragauskas et al. (2006) outline areas of research needed "to increase the impact, efficiency, and sustainability of bio-refinery facilities" (Ragauskas et al., 2006), such as research into modifying plants to enhance favorable traits, including altered cell wall structure leading to increased sugar release, as well as resilience to biotic and abiotic stress. One particular research target in *Populus* species is the decrease/alteration of the lignin content of cell walls.

A large collection of different data types has been generated for *Populus trichocarpa*. The genome has been sequenced and annotated (Tuskan et al., 2006), and the assembly is currently in its third version of revision. A collection of 1,100 accessions of *P. trichocarpa* that have been clonally propagated in four different common gardens (Tuskan et al., 2011; Slavov et al., 2012; Evans et al., 2014) have been resequenced, which has provided a large set of ∼ 28,000,000 Single Nucleotide Polymorphisms (SNPs) that has recently been publicly released (http://bioenergycenter.org/besc/gwas/). Many molecular phenotypes, such as untargetted metabolomics and pyMBMS phenotypes, that have been measured in this population provide an unparalleled resource for Genome Wide Association Studies (for example, see McKown et al., 2014). DNA methylation data in the form of MeDIP (Methyl-DNA immunoprecipitation)-seq has been performed on 10 different *P. trichocarpa* tissues (Vining et al., 2012), and gene expression has been measured across various tissues and conditions.

This study involved the development of a method to integrate these various data types in order to identify new possible candidate genes involved in target functions of interest. The importance of *P. trichocarpa* as a bioenergy crop, the availability of the high density SNP data in a GWAS population, as well as the increasing amount of genomic/phenotypic data being generated for *P. trichocarpa* made it an excellent species in which to demonstrate the method. Integrating Genome Wide Association Study (GWAS) data with other data types has previously been done to help provide context and identify relevant subnetworks/modules (Bunyavanich et al., 2014; Calabrese et al., 2017). Ritchie et al. (2015) reviewed techniques for integrating various data types for the aim of investigating gene-phenotype associations. Integrating multiple lines of evidence is a useful strategy as the more lines of evidence that connect a gene to a phenotype lowers the chance of false positives. Ritchie et al. (2015) categorized data integration approaches into two main classes, namely multi-staged analysis and meta-dimensional analysis. Multi-staged analysis aims to enrich a biological signal through various steps of analysis. Meta-dimensional analysis involves the concurrent analysis of various data types,

and is divided into three subcategories (Ritchie et al., 2015): Concatenation-based integration concatenates the data matrices of different data types into a single matrix on which a model is constructed (for example, see Fridley et al., 2012). Model-based integration involves constructing a separate model for each dataset and then constructing a final model from the results of the separate models (for example, see Kim et al., 2013). Transformation-based integration involves transforming each data type into a common form (e.g., a network) before combining them (see for example, Kim et al., 2012).

This study presents a new transformation-based integration technique: the calculation of Lines Of Evidence (LOE) scores across SNP correlation, GWAS, co-methylation, and co-expression networks for *P. trichocarpa*. Association networks for the various different data types were constructed, including a pyMBMS GWAS network, a metabolomics GWAS network, as well as co-expression, co-methylation, and SNP correlation networks, and subsequently the information in the different networks was integrated through the calculation of the newly developed Lines Of Evidence (LOE) scores. These scores quantify the number of lines of evidence connecting each gene to target functions of interest. In this work, we apply this data integration technique to the wealth of *P. trichocarpa* data in order to identify new potential genes involved in lignin biosynthesis/degradation/regulation in *P. trichocarpa*. The LOE scores represent the number of lines of evidence that exist connecting genes to lignin-related genes and phenotypes across the network layers. This is a novel multi-omic data integration approach which provides easily interpretable scores, and allows for the identification of new possible candidate genes involved in lignin biosynthesis/regulation through multiple lines of evidence. This is also the first time all of these *P. trichocarpa* datasets have been integrated on a genome-scale in a network-based manner, allowing for the easy identification of new target genes through their respective connections across network layers.
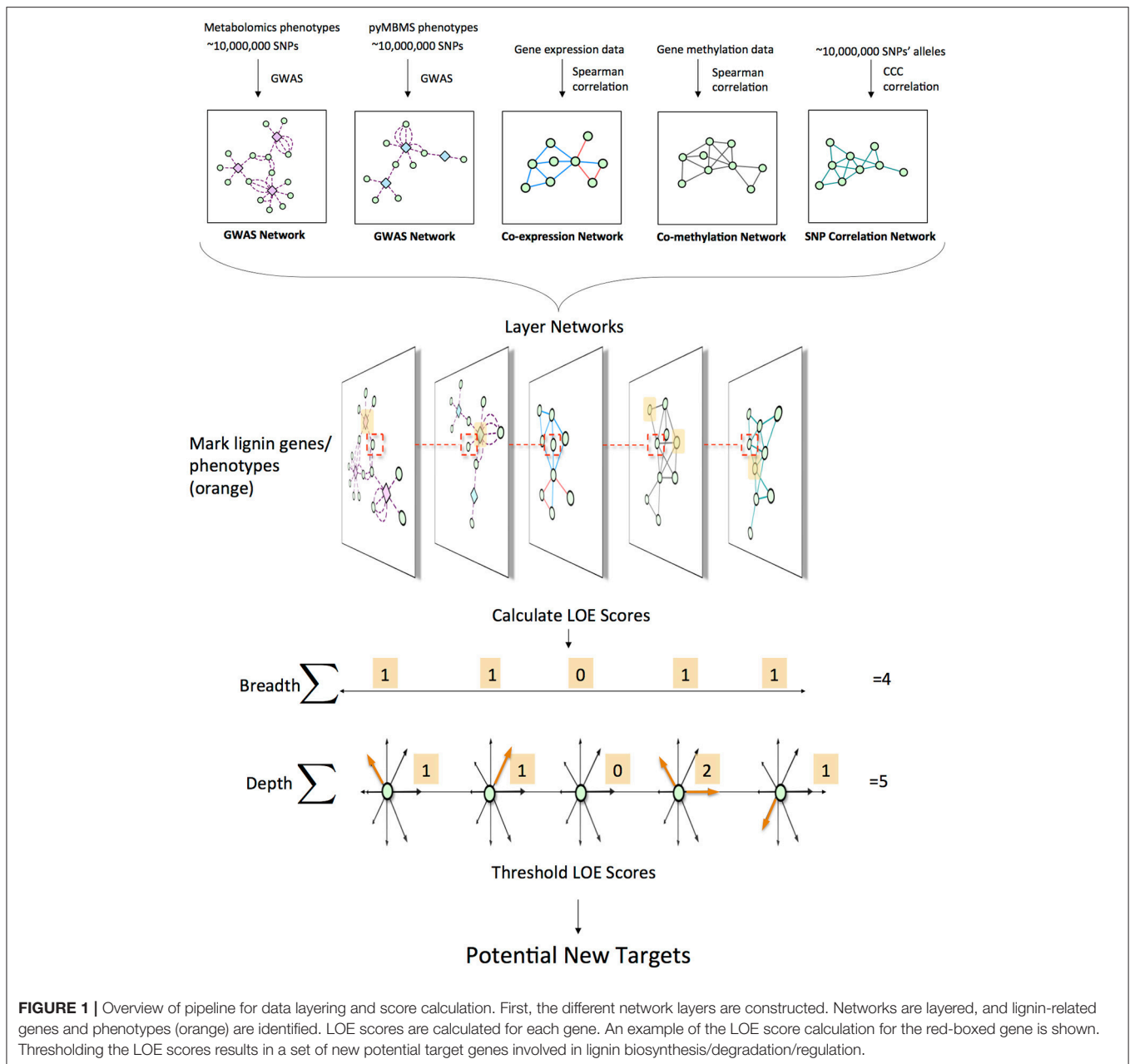
2. METHODS

2.1. Overview

This approach involved combining various data types in order to identify new possible target genes involved in lignin biosynthesis/degradation/regulation. **Figure 1** summarizes the overall approach. First, association networks were constructed including metabolomics and pyMBMS GWAS networks, co-expression, co-methylation and SNP correlation networks. Known lignin-related genes and phenotypes were then identified, and used as seeds to select lignin-related subnetworks from these various networks. The Lines Of Evidence (LOE) scoring technique was developed, and each gene was then scored based on its Lines Of Evidence linking it to lignin-related genes and phenotypes.

2.2. Metabolomics Phenotype Data

The *P. trichocarpa* leaf samples for 851 unique clones were collected over 3 consecutive sunny days in July 2012. For 200 of those clones, a second biological replicate was also sampled. Typically, leaves (leaf plastocron index 9 plus or minus 1) on a

---

Abbreviations: GWAS, Genome Wide Association Study; LOE, Lines of Evidence; SNP, Single Nucleotide Polymorphism; pyMBMS, Pyrolysis Molecular Beam Mass Spectrometer; DNA, Deoxyribonucleic acid; MeDIP, Methyl-DNA immunoprecipitation; TPM, Transcripts Per Million; CCC, Custom Correlation Coefficient; PNT, Potential New Targets.

FIGURE 1 | Overview of pipeline for data layering and score calculation. First, the different network layers are constructed. Networks are layered, and lignin-related genes and phenotypes (orange) are identified. LOE scores are calculated for each gene. An example of the LOE score calculation for the red-boxed gene is shown. Thresholding the LOE scores results in a set of new potential target genes involved in lignin biosynthesis/degradation/regulation.

south facing branch from the upper canopy of each tree were quickly collected, wiped with a wet tissue to clean both surfaces and the leaf then fast frozen under dry ice. Leaves were kept on dry ice and shipped back to the lab and stored at $-80°C$ until processed for analyses. Metabolites from leaf samples were lyophilized and then ground in a micro-Wiley mill (1 mm mesh size). Approximately 25 mg of each sample was twice extracted in 2.5 mL 80% ethanol (aqueous) for 24 h with the extracts combined, and 0.5 mL dried in a helium stream. "Sorbitol [(75 μl of a 1 mg/mL aqueous solution)] was added ... before extraction as an internal standard to correct for differences in extraction efficiency, subsequent differences in derivatization efficiency and changes in sample volume during heating" (Zhao et al.,

2015). Metabolites in the dried sample extracts were converted to their trimethylsilyl (TMS) derivatives, and analyzed by gas chromatography-mass spectrometry, as described previously (Li et al., 2012; Tschaplinski et al., 2012), and also described here: Timm et al. (2016): Briefly, dried extracts of metabolites "were dissolved in acetonitrile followed by the addition of N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS), and samples then heated for 1 h at 70°C to generate trimethylsilyl (TMS) derivatives (Li et al., 2012; Tschaplinski et al., 2012). After 2 days, aliquots were injected into an Agilent 5975C inert XL gas chromatograph-mass spectrometer (GC-MS). The standard quadrupole GC-MS is operated in the electron impact (70 eV) ionization mode,

targeting 2.5 full-spectrum (50–650 Da) scans per second, as described previously (Tschaplinski et al., 2012). Metabolite peaks were extracted using a key selected ion, characteristic m/z fragment, rather than the total ion chromatogram, to minimize integrating co-eluting metabolites" (quotation from Timm et al., 2016). As described in Zhao et al. (2015): "...[The peak areas were] normalized to the quantity of the internal standard (sorbitol) [injected, and the] amount of sample extracted... A large user-created database (>2,400 spectra) of mass spectral electron impact ionization (EI) fragmentation patterns of TMS-derivatized metabolites, as well as the Wiley Registry [10th] Edition combined with NIST [2014] mass spectral database, were used to identify the metabolites of interest to be quantified" (Zhao et al., 2015) (Brackets indicate deviations from quoted text).
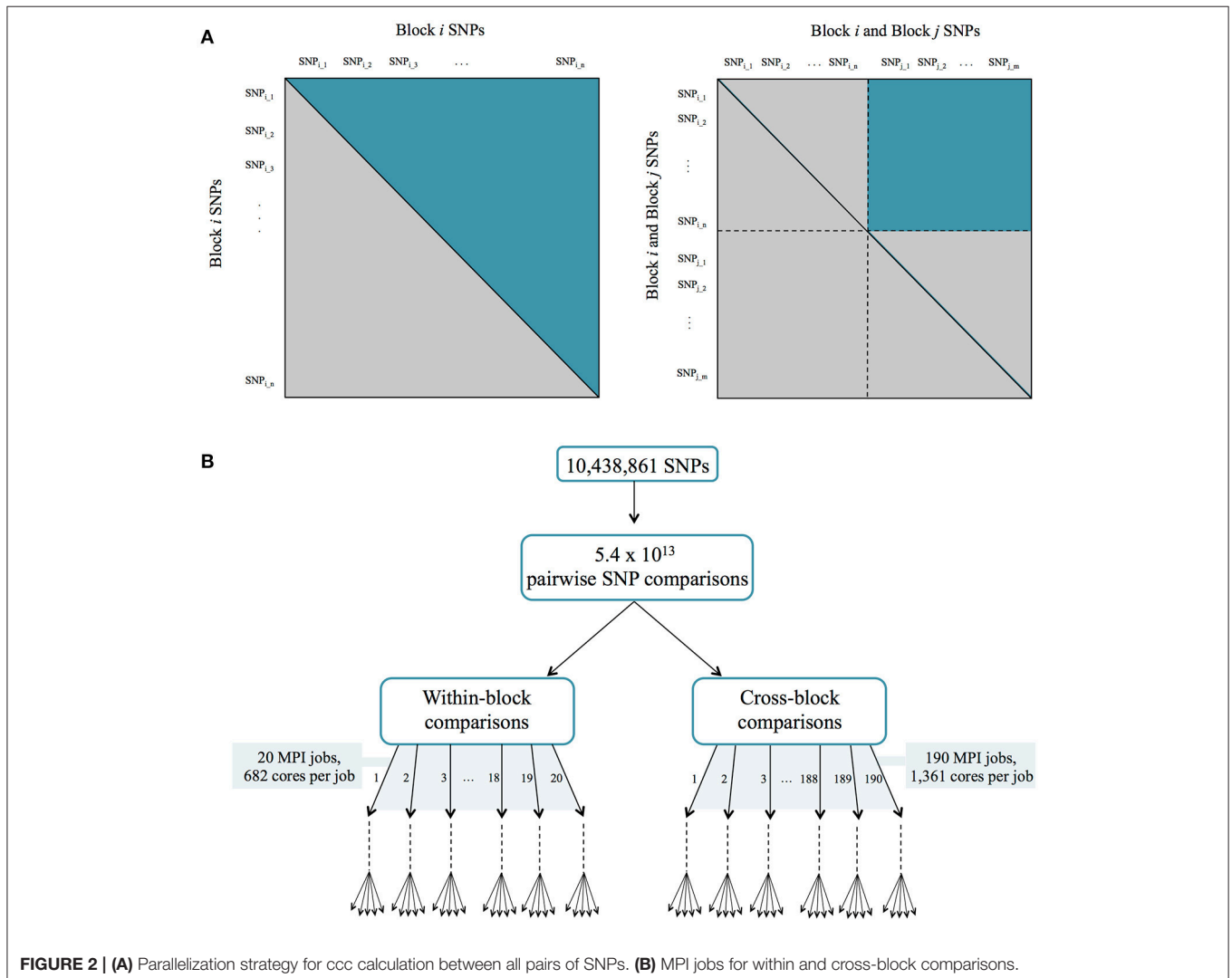
2.3. pyMBMS Phenotype Data

The pyMBMS phenotype data was generated using the method as described in Biswal et al. (2015): "A commercially

available molecular beam mass spectrometer (MBMS) designed specifically for biomass analysis was used for pyrolysis vapor analysis (Evans and Milne, 1987; Tuskan et al., 1999; Sykes et al., 2009). Approximately 4 mg of air dried 20 mesh biomass was introduced into the quartz pyrolysis reactor via 80 μL deactivated stainless steel Eco-Cups provided with the autosampler. Mass spectral data from m/z 30–450 were acquired on a Merlin Automation data system version 3.0 using 17 eV electron impact ionization."

The pyMBMS mz peaks were annotated as described in Sykes et al. (2009), as done previously in Muchero et al. (2015).

2.4. Single Nucleotide Polymorphism Data

A dataset consisting of 28,342,758 SNPs called across 882 *P. trichocarpa* (Tuskan et al., 2006) genotypes was obtained from http://bioenergycenter.org/besc/gwas/. This dataset is derived from whole genome sequencing of undomesticated *P. trichocarpa* genotypes collected from the U.S. and Canada, and clonally replicated in common gardens (Tuskan et al.,

FIGURE 2 | (A) Parallelization strategy for ccc calculation between all pairs of SNPs. **(B)** MPI jobs for within and cross-block comparisons.
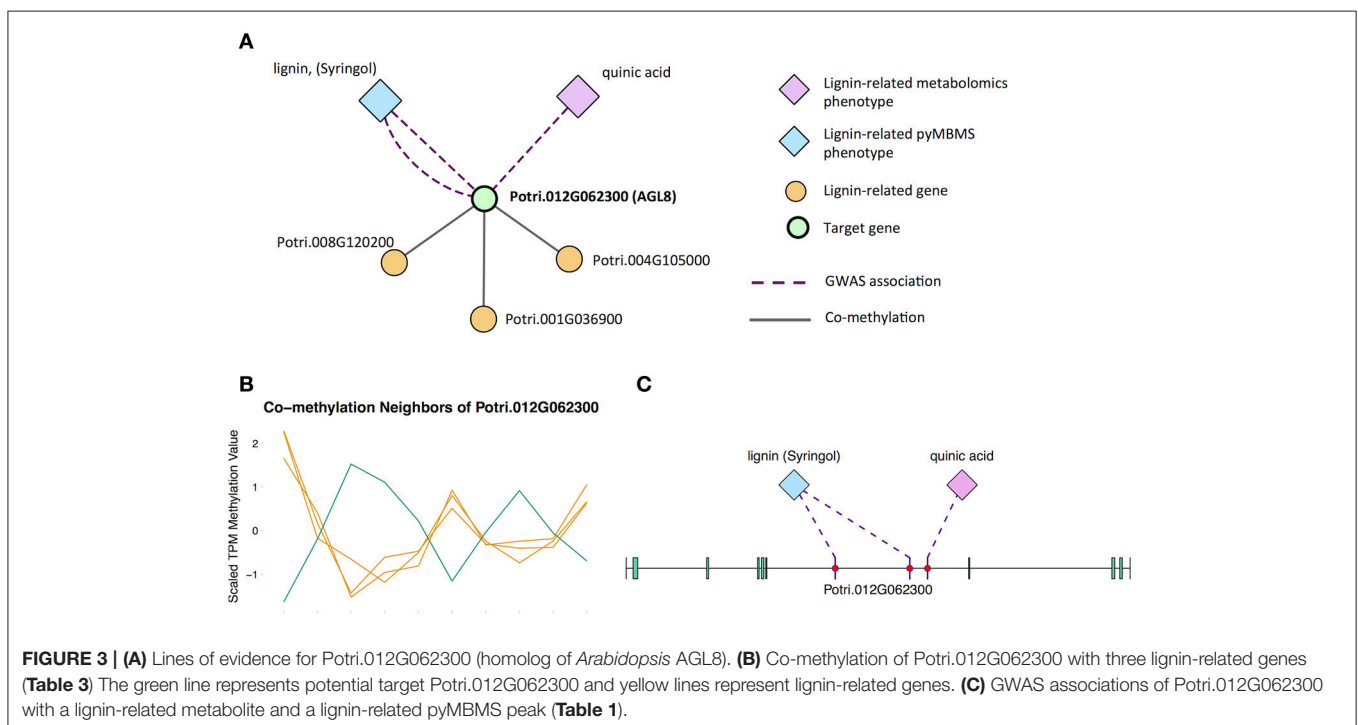
2011). Genotypes from this population have previously been used for population genomics (Evans et al., 2014) and GWAS studies in *P. trichocarpa* (McKown et al., 2014) as well as for investigating linkage disequilibrium in the population (Slavov et al., 2012).

Whole genome resequencing was carried out on a sample 882 *P. trichocarpa* natural individuals to an expected median coverage of 15x using Illumina Genome Analyzer, HiSeq 2000, and HiSeq 2500 sequencing platforms at the DOE Joint Genome Institute. Alignments to the *P. trichocarpa* Nisqually-1 v.3.0 reference genome were performed using BWA v0.5.9-r16 with default parameters, followed by post-processing with the picard FixMateInformation and MarkDuplicates tools. Genetic variants were called by means of the Genome Analysis Toolkit v. 3.5.0 (GATK; Broad Institute, Cambridge, MA, USA) (McKenna et al., 2010; van der Auwera et al., 2013). Briefly, variants were called independently for each individual using the concatenation of RealignerTargetCreator, IndelRealigner, and HaplotypeCaller tools, and the whole population was combined using GenotypeGVCFs, obtaining a dataset with all the variants detected across the sample population. Biallelic SNPs were extracted using the SelectVariants tool and quality-filtered using the GATK's machine-learning implementation Variant Quality Score Recalibration (VQSR). To this end, the tool VariantRecalibrator was used to create the recalibration file and the sensitivity tranches file. As a "truth" dataset, we used SNP calls from a population of seven female and seven male *P. trichocarpa* that had been crossed in a half diallel design. "True" SNPs were identified by the virtual absence of segregation distortion and Mendelian violations in the progeny of these 49 crosses (ca. 500

offspring in total). As a "non-true" dataset, we used the SNP calls of seven open-pollinated crosses from these 7 females ($n = 90$), filtered using hard-filtering methods recommended in the GATK documentation (tool: VariantFiltration; quality thresholds: QD $<$ 1.5, FS $>$ 75.0, MQ $<$ 35.0, missing alleles $<$ 0.5 and MAF $>$ 0.05). The prior likelihoods for the true and non-true datasets were $Q = 15$ and $Q = 10$, respectively, and the variant quality annotations to define the variant recalibration space were DP, QD, MQ, MQRankSum, ReadPosRankSum, FS, SOR, and InbreedingCoeff. Finally, we used the ApplyRecalibration tool on the full GWAS dataset to assign SNPs to tranches representing different levels of confidence. We selected SNPs in the tranche with true sensitivity $<$90, which minimizes false positives, but at an expected cost of 10% false negatives. The final filtered dataset had a transition/transversion ratio of 2.07, compared to 1.88 for the unfiltered SNPs. To further validate the quality of these SNP calls, we compared them to an Illumina Infinium BeadArray that had been generated from a subset of this population dataset (Geraldes et al., 2013). The average match rate was 96% ($\pm$2% SD) for 641 individuals across 20,723 loci.

SNPs in this dataset were divided into different Tranches, indicating the percentage of "true" SNPs recovered. For further analysis in this study, we made use of the PASS SNPs, corresponding to the most stringent Tranche, recovering 90% of the true SNPs (see http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr). VCFtools (Danecek et al., 2011) was used to extract the desired Tranche of SNPs from the VCF file and reformat it into .tfam and .tped files.

FIGURE 3 | (A) Lines of evidence for Potri.012G062300 (homolog of *Arabidopsis* AGL8). **(B)** Co-methylation of Potri.012G062300 with three lignin-related genes (**Table 3**) The green line represents potential target Potri.012G062300 and yellow lines represent lignin-related genes. **(C)** GWAS associations of Potri.012G062300 with a lignin-related metabolite and a lignin-related pyMBMS peak (**Table 1**).

2.5. GWAS Network Construction

The metabolomics and pyMBMS data was used as phenotypes in a genome wide association analysis. The respective phenotype measured over all the genotypes were analyzed to account for potential outliers. A median absolute deviation (MAD) from the median (Leys et al., 2013) cutoff was applied to determine if a particular measurement of a given phenotype was an outlier with respect to all measurements of that phenotype across the population. To account for asymmetry, the deviation values were estimated separately for values below and above the median, respectively. The distribution of the measured values together with the distribution of their estimated deviation was analyzed and a cutoff of 5 was determined to identify putative outlier values. Phenotypes that had non-outlier measurements in at least 20 percent of the population were retained for further analysis, this was to ensure sufficient signal for the genome wide association model. This resulted in 1,262

pyMBMS derived phenotypes and 818 metabolomics derived phenotypes.

To estimate the statistical significant associations between the respective phenotypes and the SNPs called across the population, we applied a linear mixed model using EMMAX (Kang et al., 2010). Taking into account population structure estimated from a kinship matrix, we tested each of the respective 2,080 phenotypes against the high-confidence SNPs and corrected for multiple hypotheses bias using the Benjamini–Hochberg control for false-discovery rate of 0.1 (Benjamini and Hochberg, 1995). This was done in parallel with a Python wrapper that utilized the schwimmbad Python package (Price-Whelan and Foreman-Mackey, 2017).

SNP-Phenotype GWAS networks were then pruned to only include SNPs that resided within genes, and SNPs were mapped to their respective genes, resulting in a gene-phenotype network. SNPs were determined to be

TABLE 1 | GWAS associations for select new potential target genes, indicating the SNP(s) within the potential new target gene which are associated with the lignin-related phenotype(s).

Source SNP	Source gene	Target phenotype
GWAS Associations for Potri.012G062300 (AGL8, AT5G60910)		
12:6952245	Potri.012G062300	Quinic acid
12:6948543	Potri.012G062300	Lignin (Syringol)
12:6951532	Potri.012G062300	Lignin (Syringol)
GWAS Associations for Potri.013G102600 (AGL12, AT1G71692)		
13:11604094	Potri.013G102600	3-O-caffeoyl-quinate
13:11606331	Potri.013G102600	Coumaroyl-tremuloidin
13:11600422	Potri.013G102600	Coumaroyl-tremuloidin
13:11601236	Potri.013G102600	Hydroxyphenyl lignan glycoside
GWAS Associations for Potri.007G115100 (AGL22, AT2G22540/AGL24, AT4G24540)		
07:13650194	Potri.007G115100	Caffeoyl conjugate
07:13651354	Potri.007G115100	Caffeoyl conjugate
07:13642539	Potri.007G115100	Caffeoyl conjugate
07:13639923	Potri.007G115100	Lignin, syringyl (Syringaldehyde)
GWAS Associations for Potri.009G053900 (MYB46, AT5G12870)		
09:5768381	Potri.009G053900	Hydroxyphenyl lignan glycoside
GWAS Associations for Potri.010G141000 (MYB111, AT5G49330)		
10:15273000	Potri.010G141000	Benzoyl-salicylate caffeic acid conjugate
GWAS Associations for Potri.006G170800 (MYB36, AT5G57620)		
06:17847162	Potri.006G170800	mz 297, RT 17.14
GWAS Associations for Potri.016G078600 (CPSRP54, AT5G03940)		
16:5995136	Potri.016G078600	Caffeoyl conjugate
16:5995136	Potri.016G078600	Feruloyl conjugate
16:5996083	Potri.016G078600	Salicyl-coumaroyl-glucoside
16:5999408	Potri.016G078600	Salicyl-coumaroyl-glucoside
16:5999474	Potri.016G078600	Salicyl-coumaroyl-glucoside
16:6000236	Potri.016G078600	Salicyl-coumaroyl-glucoside

Additional RT and mz information for partially identified metabolites can be seen in Table S3.

within genes using the gene boundaries defined in the `Ptrichocarpa_210_v3.0.gene.gff3` from the *P. trichocarpa* version 3.0 genome assembly on Phytozome (Goodstein et al., 2012).
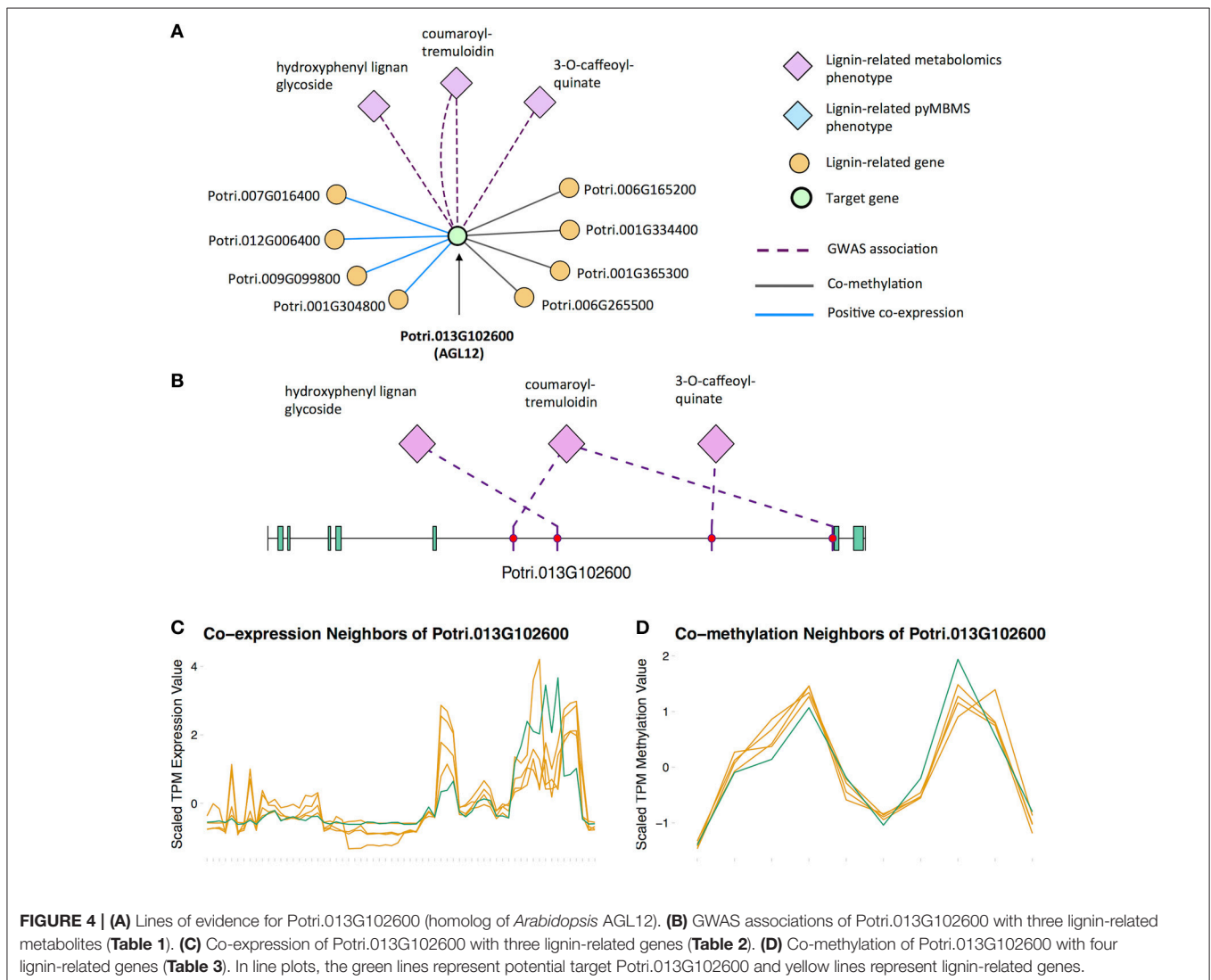
2.6. Gene Expression Data

P. trichocarpa (Nisqually-1) RNA-seq dataset from JGI Plant Gene Atlas project (Sreedasyam et al., unpublished) was obtained from Phytozome. This dataset consists of samples for standard tissues (leaf, stem, root, and bud tissue) and libraries generated from nitrogen source study. List of sample descriptions was accessed from: https://phytozome.jgi.doe.gov/phytomine/aspect. do?name=Expression.

P. trichocarpa (Nisqually-1) cuttings were potted in 4″ X 4″ X 5″ containers containing 1:1 mix of peat and perlite. Plants were grown under 16-h-light/8-h-dark conditions, maintained at 20–23 °C and an average of 235 μmol m^{-2}s^{-1} to generate tissue for (1) standard tissues and (2) nitrogen source study. Plants for standard tissue experiment were watered with McCown's

woody plant nutrient solution and plants for nitrogen experiment were supplemented with either 10 mM KNO3 (NO3$^-$ plants) or 10mM NH4Cl (NH4+ plants) or 10 mM urea (urea plants). Once plants reached leaf plastochron index 15 (LPI-15), leaf, stem, root, and bud tissues were harvested and immediately flash frozen in liquid nitrogen and stored at −80°C until further processing was done. Every harvest involved at least three independent biological replicates for each condition and a biological replicate consisted of tissue pooled from 3 plants.

RNA extraction and sequencing was performed as previously described in McCormick et al. (2018). Tissue was ground under liquid nitrogen and high quality RNA was extracted using standard Trizol-reagent based extraction (Li and Trick, 2005). The integrity and concentration of the RNA preparations were checked initially using Nano-Drop ND-1000 (Nano-Drop Technologies) and then by BioAnalyzer (Agilent Technologies). "Plate-based RNA sample prep was performed on the PerkinElmer Sciclone NGS robotic liquid handling system

FIGURE 4 | (A) Lines of evidence for Potri.013G102600 (homolog of *Arabidopsis* AGL12). **(B)** GWAS associations of Potri.013G102600 with three lignin-related metabolites (**Table 1**). **(C)** Co-expression of Potri.013G102600 with three lignin-related genes (**Table 2**). **(D)** Co-methylation of Potri.013G102600 with four lignin-related genes (**Table 3**). In line plots, the green lines represent potential target Potri.013G102600 and yellow lines represent lignin-related genes.

using Illumina's TruSeq Stranded mRNA HT sample prep kit utilizing poly-A selection of mRNA following the protocol outlined by Illumina in their user guide: http://support.illumina. com/sequencing/sequencing_kits/truseq_stranded_mrna_ht_ sample_prep_kit.html, v. The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v4, and Illumina's cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcell was performed on the Illumina HiSeq2500 sequencer using HiSeq TruSeq SBS sequencing kits, v4, following a 2 × 150 indexed run recipe" (McCormick et al., 2018).

2.7. Co-expression Network Construction

Gene expression atlas data for *P. trichocarpa* consisting of 63 different samples were used to construct a co-expression network. Reads were trimmed using Skewer (Jiang et al., 2014). Star (Dobin et al., 2013) was then used to align the reads to the *P. trichocarpa* reference genome (Tuskan et al., 2006) obtained from Phytozome (Goodstein et al., 2012). TPM (Transcripts Per Million) expression values (Wagner et al., 2012) were then calculated for each gene. This resulted in a gene expression matrix *E* in which rows represented genes, columns represented samples and each entry *ij* represented the expression (TPM) of gene *i* in sample *j*. The Spearman correlation coefficient was then calculated between the expression profiles of all pairs of genes (i.e. all pairs of rows of the matrix *E*) using the mcxarray and mcxdump programs from the MCL-edge package (van Dongen, 2000, 2008) available from http://micans.org/mcl/. This was performed in parallel using Perl wrappers making use of the

Parallel::MPI::Simple Perl module, (Alex Gough, http://search. cpan.org/~ajgough/Parallel-MPI-Simple-0.03/Simple.pm) using compute resources at the Oak Ridge Leadership Computing Facility (OLCF).

Figure S1A shows the distribution of Spearman correlation values for the co-expression network. An absolute threshold of 0.85 was applied.

2.8. Co-methylation Network Construction

Methylation data for *P. trichocarpa* (Vining et al., 2012) re-aligned to the version 3.0 assembly of *P. trichocarpa* was obtained from Phytozome (Goodstein et al., 2012). This data consisted of MeDIP-seq (Methyl-DNA immunoprecipitation-seq) reads from 10 different *P. trichocarpa* tissues, including bud, callus, female catkin, internode explant, leaf, male catkin, phloem, regenerated internode, root, and xylem tissue.

BamTools stats (Barnett et al., 2011) was used to determine basic properties of the reads in each .bam file. Samtools (Li et al., 2009) was then used to extract only mapped reads. The number of reads which mapped to each gene feature was determined using htseq-count (Anders et al., 2014). These read counts were then converted to TPM values (Wagner et al., 2012), providing a methylation score for each gene in each tissue. The TPM value for a gene *g* in a given sample was defined as:

$$TPM_g = \frac{\frac{c_g}{l_g} \times 10^6}{\sum_g \frac{c_g}{l_g}} \quad (1)$$

where c_g is the number of reads mapped to gene *g* and l_g is the length of gene *g* in kb, calculated by subtracting

TABLE 2 | Co-expression associations for select new potential target genes. Annotations are derived from best *Arabidopsis* hit descriptions, GO terms and in some cases MapMan annotations.

Source gene	Target gene	Target *Arabidopsis* best hit	Annotation
Co-expression Associations for Potri.013G102600 (AGL12, AT1G71692)			
Potri.013G102600	Potri.001G304800	AT4G34050	Caffeoyl Coenzyme A O-Methyltransferase 1
Potri.013G102600	Potri.009G099800	AT4G34050	Caffeoyl Coenzyme A O-Methyltransferase 1
Potri.013G102600	Potri.012G006400	AT5G54160	Caffeate O-Methyltransferase 1
Potri.013G102600	Potri.007G016400	AT4G36220	Ferulic acid 5-hydroxylase 1
Co-expression Associations for Potri.009G053900 (MYB46, AT5G12870)			
Potri.009G053900	Potri.003G100200	AT1G32100	Pinoresinol reductase 1
Potri.009G053900	Potri.012G006400	AT5G54160	Caffeate O-Methyltransferase 1
Co-expression Associations for Potri.010G141000 (MYB111, AT5G49330)			
Potri.010G141000	Potri.007G030300	AT3G50740	UDP-glucosyl transferase 72E1
Co-expression Associations for Potri.006G170800 (MYB36, AT5G57620)			
Potri.006G170800	Potri.001G362800	AT3G26300	Cytochrome P450, family 71, subfamily B, polypeptide 34/F5H
Potri.006G170800	Potri.016G106100	AT3G09220	laccase 7
Potri.006G170800	Potri.013G120900	AT4G35160	N-acetylserotonin O-methyltransferase
Co-expression Associations for Potri.016G078600 (CPSRP54, AT5G03940)			
Potri.016G078600	Potri.003G096600	AT2G35500	Shikimate kinase like 2
Potri.016G078600	Potri.017G062800	AT3G26900	Shikimate kinase like 1

the gene start position from the gene end position, and dividing the resulting difference by 1,000. A methylation matrix M was then formed, in which rows represented genes, columns represented tissues and each entry ij represented the methylation score (TPM) of gene i in tissue j. A co-methylation network (see Davies et al., 2012; Akulenko and Helms, 2013; Busch et al., 2016) was then constructed by calculating the Spearman correlation coefficient between the methylation profiles of all pairs of genes using mcxarray and mcxdump programs from the MCL-edge package (van Dongen, 2000, 2008) http://micans.org/mcl/. Figure S1B shows the distribution of Spearman Correlation values. An absolute threshold of 0.95 was applied.

Read counting using htseq-count, as well as Spearman correlation calculations were performed in parallel using Perl wrappers making use of the Parallel::MPI::Simple Perl module, developed by Alex Gough and available on The Comprehensive Perl Archive Network (CPAN) at www.cpan.org and used compute resources at the Oak Ridge Leadership Computing Facility (OLCF).

2.9. SNP Correlation Network Construction

The Custom Correlation Coefficient (CCC) (Climer et al., 2014a,b) was used to calculate the correlation between the occurrence of pairs of SNPs across the 882 genotypes. The CCC between allele x at position i and allele y and position j is defined as:

$$CCC_{i_x j_y} = \frac{9}{2} R_{i_x j_y} \left(1 - \frac{1}{f_{i_x}} \right) \left(1 - \frac{1}{f_{j_y}} \right) \qquad (2)$$

where $R_{i_x j_y}$ is the relative co-occurrence of allele x at position i and allele y at position j, f_{i_x} is the frequency of allele x at position i and f_{j_y} is the frequency of allele y at position j.

This was performed in a parallel fashion using similar computational approaches as described for the co-expression network above. The set of ~10 million SNPs was divided into 20 different blocks, and the CCC was calculated for each within-block and cross-block SNPs in separate jobs, to a total of 210 MPI jobs (**Figure 2**). A threshold of 0.7 was then applied. The resulting SNP correlation network was pruned to only include SNPs that resided within genes and SNPs were mapped to the

TABLE 3 | Co-methylation associations for select new potential target genes. Annotations are derived from best *Arabidopsis* hit descriptions and GO terms and in some cases MapMan annotations.

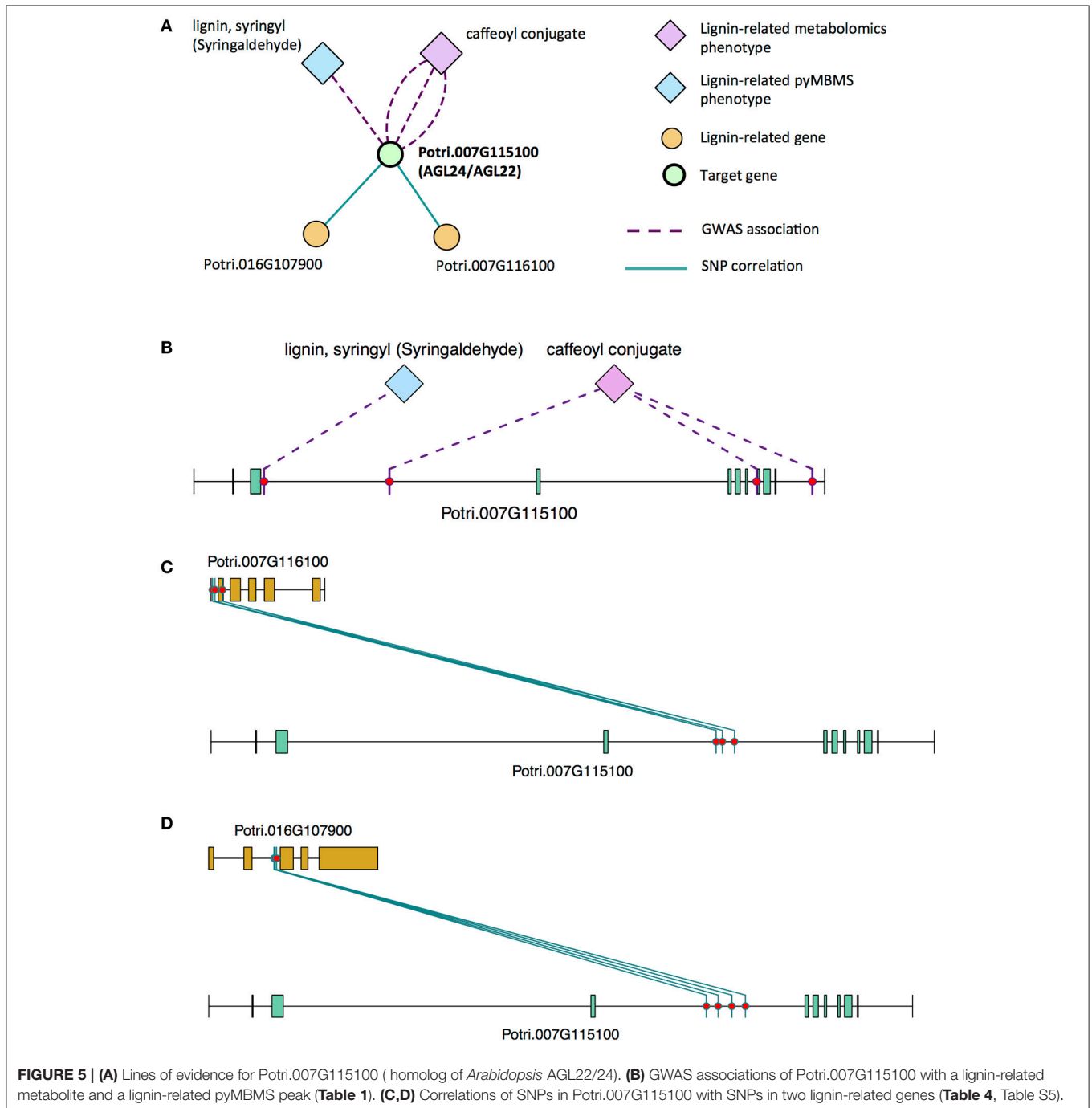| Source gene	Target gene	Target *Arabidopsis* best hit	Annotation
Co-methylation Associations for Potri.012G062300 (AGL8, AT5G60910)			
Potri.012G062300	Potri.001G036900	AT3G21240	4-coumarate:CoA ligase 2
Potri.012G062300	Potri.008G120200	AT1G68540	Cinnamoyl CoA reductase-like 6
Potri.012G062300	Potri.004G105000	AT5G14700	(NAD(P)-binding Rossmann-fold Superfamily protein, cinnamoyl-CoA reductase activity/CCR1
Co-methylation Associations for Potri.013G102600 (AGL12, AT1G71692)			
Potri.013G102600	Potri.001G334400	AT5G63380	4-coumarate-CoA ligase activity /4CL
Potri.013G102600	Potri.001G365300	AT3G26300	cytochrome P450, family 71, subfamily B, polypeptide 34/F5H
Potri.013G102600	Potri.006G265500	AT5G10820	Major facilitator superfamily protein/Phenylpropanoid pathway
Potri.013G102600	Potri.006G165200	AT2G19070	Spermidine hydroxycinnamoyl transferase
Co-methylation Associations for Potri.009G053900 (MYB46, AT5G12870)			
Potri.009G053900	Potri.008G196100	AT3G06350	Bi-functional dehydroquinate-shikimate dehydrogenase enzyme
Potri.009G053900	Potri.002G018300	AT4G39330	Cinnamyl alcohol dehydrogenase 9
Potri.009G053900	Potri.004G102000	AT4G05160	4-coumarate-CoA ligase activity/4CL)
Potri.009G053900	Potri.008G136600	AT1G67980	Caffeoyl-CoA 3-O-methyltransferase
Co-methylation Associations for Potri.010G141000 (MYB111, AT5G49330)			
Potri.010G141000	Potri.008G196100	AT3G06350	Bi-functional dehydroquinate-shikimate dehydrogenase enzyme
Potri.010G141000	Potri.004G102000	AT4G05160	4-coumarate-CoA ligase activity/4CL
Potri.010G141000	Potri.008G074500	AT5G34930	Arogenate dehydrogenase
Potri.010G141000	Potri.005G028000	AT5G48930	Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase
Potri.010G141000	Potri.018G100500	AT2G23910	NAD(P)-binding Rossmann-fold superfamily protein, cinnamoyl-CoA reductase activity/CCR1
Potri.010G141000	Potri.010G230200	AT1G20510	OPC-8:0 CoA ligase1, 4-coumarate-CoA ligase activity/4CL
Co-methylation Associations for Potri.006G170800 (MYB36, AT5G57620)			
Potri.006G170800	Potri.016G093700	AT4G05160	AMP-dependent synthetase and ligase family, 4-coumarate-CoA ligase activity/4CL
Co-methylation Associations for Potri.016G078600 (CPSRP54, AT5G03940)			
Potri.016G078600	Potri.014G135500	AT3G06350	Bi-functional dehydroquinate-shikimate dehydrogenase enzyme

genes in which they reside. Gene boundaries used were defined in the `Ptrichocarpa_210_v3.0.gene.gff3` file from the *P. trichocarpa* version 3.0 genome assembly on Phytozome (Goodstein et al., 2012). A local LD filter was then set, retaining correlations between SNPs greater than 10kb apart. The distribution of CCC values can be seen in Figure S1C (Supplementary Note 1).

2.10. Target Lignin Genes/Phenotypes

A scoring system was developed in order to quantify the Lines Of Evidence (LOE) linking each gene to lignin-related genes/phenotypes. The LOE scores quantify the number of lines linking each gene to lignin-related genes and phenotypes across the different network data layers. Thus, the method requires as input a list of known lignin-related genes/phenotypes.

P. trichocarpa gene annotations in the `Ptrichocarpa_210_v3.0.annotation_info.txt` file from the version 3.0 genome assembly were used, available on Phytozome (Goodstein et al., 2012). This included *Arabidopsis* best hits and corresponding gene descriptions, as well as GO terms (Ashburner et al.,

FIGURE 5 | (A) Lines of evidence for Potri.007G115100 (homolog of *Arabidopsis* AGL22/24). (B) GWAS associations of Potri.007G115100 with a lignin-related metabolite and a lignin-related pyMBMS peak (**Table 1**). (C,D) Correlations of SNPs in Potri.007G115100 with SNPs in two lignin-related genes (**Table 4**, Table S5).

2000; Gene Ontology Consortium, 2017) and Pfam domains (Finn et al., 2016). Genes were also assigned MapMan annotations using the Mercator tool (Lohse et al., 2014).

Lignin building blocks (monolignols) are derived from phenylalanine in the phenylpropanoid and monolignol pathways, and phenylalanine itself is produced from the shikimate pathway (Vanholme et al., 2010). To compile a list of *P. trichocarpa* genes which are related to the biosynthesis of lignin, *P. trichocarpa* genes were assigned MapMan annotations using the Mercator tool (Lohse et al., 2014). Genes in the Shikimate (MapMan bins 13.1.6.1, 13.1.6.3, and 13.1.6.4), Phenylpropanoid (MapMan bin 16.2) and Lignin/Lignan (MapMan bin 16.2.1) pathways were then selected. A list of these lignin-related genes and their MapMan annotations can be seen in Table S1.

Lignin-related pyMBMS peaks, as described in Sykes et al. (2009), Davis et al. (2006), and Muchero et al. (2015) were identified among the pyMBMS GWAS hits, and are shown in Table S2. Lignin-related metabolites and metabolites in the lignin pathway were also identified among the metabolomics GWAS hits, a list of which can be seen in Table S3. For partially identified metabolites, additional RT and mz information can be seen in Table S3.

2.11. Extraction of Lignin-Related Subnetworks

Let L_G, L_M, and L_P represent our sets of lignin-related genes, metabolites and pyMBMS peaks, respectively (Tables S1–S3). A network can be defined as $N = (V, E)$ where V is the set of nodes and E is the set of edges connecting nodes in V. In particular, let the co-expression network be represented by $N_{coex} = (V_{coex}, E_{coex})$, the co-methylation network by $N_{cometh} = (V_{cometh}, E_{cometh})$ and the SNP correlation network by $N_{snp} = (V_{snp}, E_{snp})$. The GWAS networks can be represented as bipartite networks $N = (U, V, E)$ where U is the set of phenotype nodes, V is the set of gene nodes, and E is the set of edges, with each edge e_{ij} connecting node $i \in U$ with node $j \in V$. Let the metabolomics GWAS network be represented by $N_{metab} = (U_{metab}, V_{metab}, E_{metab})$ and the pyMBMS GWAS network by $N_{pymbms} = (U_{pymbms}, V_{pymbms}, E_{pymbms})$. We construct the *guilt by association* subnetworks of genes connected to lignin-related genes/phenotypes as follows:

N_{coex}^L is the subnetwork of N_{coex} including the lignin related genes $l \in L_G$ and their direct neighbors:

$$N_{coex}^L = (V_{coex}^L, E_{coex}^L) \text{ where} \tag{3}$$

$$V_{coex}^L = \{g | g \in (L_G \cap V_{coex})\} \cup \{g | (g \in V_{coex})$$
$$\wedge (\exists l \in L_G | \{l, g\} \in E_{coex})\} \tag{4}$$

$$E_{coex}^L = \{e = \{i, j\} \in E_{coex} | i \in V_{coex}^L \wedge j \in V_{coex}^L\} \tag{5}$$

N_{cometh}^L is the subnetwork of N_{cometh} including the lignin related genes $l \in L_G$ and their direct neighbors:

$$N_{cometh}^L = (V_{cometh}^L, E_{cometh}^L) \text{ where} \tag{6}$$

$$V_{cometh}^L = \{g | g \in (L_G \cap V_{cometh})\} \cup \{g | (g \in V_{cometh})$$
$$\wedge (\exists l \in L_G | \{l, g\} \in E_{cometh})\} \tag{7}$$

$$E_{cometh}^L = \{e = \{i, j\} \in E_{cometh} | i \in V_{cometh}^L \wedge j \in V_{cometh}^L\} \tag{8}$$

N_{snp}^L is the subnetwork of N_{snp} including the lignin related genes $l \in L_G$ and their direct neighbors:

$$N_{snp}^L = (V_{snp}^L, E_{snp}^L) \text{ where} \tag{9}$$

$$V_{snp}^L = \{g | g \in (L_G \cap V_{snp})\} \cup \{g | (g \in V_{snp})$$
$$\wedge (\exists l \in L_G | \{l, g\} \in E_{snp})\} \tag{10}$$

$$E_{snp}^L = \{e = \{i, j\} \in E_{snp} | i \in V_{snp}^L \wedge j \in V_{snp}^L\} \tag{11}$$

N_{metab}^L is the subnetwork of N_{metab} including the lignin related metabolites $m \in L_M$ and their direct neighboring genes:

$$N_{metab}^L = (U_{metab}^L, V_{metab}^L, E_{metab}^L) \text{ where} \tag{12}$$

$$U_{metab}^L = \{m | m \in (L_M \cap U_{metab})\} \tag{13}$$

$$V_{metab}^L = \{g | (g \in V_{metab}) \wedge (\exists m \in L_M | (m, g) \in E_{metab})\} \tag{14}$$

$$E_{metab}^L = \{e = (i, j) \in E_{metab} | i \in U_{metab}^L \wedge j \in V_{metab}^L\} \tag{15}$$

TABLE 4 | SNP correlation associations for select new potential target genes. Annotations are derived from best *Arabidopsis* hit descriptions, GO terms and in some cases MapMan annotations.

Source gene	Target gene	Target *Arabidopsis* best hit	Annotation
SNP Correlations for Potri.007G115100 (AGL22, AT2G22540/AGL24, AT4G24540)			
Potri.007G115100	Potri.007G116100	AT2G22570	Nicotinamidase 1
Potri.007G115100	Potri.016G107900	AT3G09220	Laccase 7
SNP Correlations for Potri.016G078600 (CPSRP54, AT5G03940)			
Potri.016G078600	Potri.016G078300	AT4G37970	Cinnamyl alcohol dehydrogenase 6

N^L_{pymbms} is the subnetwork of N_{pymbms} including the lignin related pyMBMS peaks $p \in L_P$ and their direct neighboring genes:

$$N^L_{pymbms} = (U^L_{pymbms}, V^L_{pymbms}, E^L_{pymbms}) \text{ where} \qquad (16)$$

$$U^L_{pymbms} = \{p | p \in (L_P \cap U_{pymbms})\} \qquad (17)$$

$$V^L_{pymbms} = \{g | (g \in V_{pymbms}) \land (\exists p \in L_P | (p,g) \in E_{pymbms})\} \qquad (18)$$

$$E^L_{pymbms} = \{e = (i,j) \in E_{pymbms} | i \in U^L_{pymbms} \land j \in V^L_{pymbms}\} \qquad (19)$$

2.12. Calculating LOE Scores

For a given gene g, the *degree* of that gene $D(g)$ indicates the number of connections that the gene has in a given network. Let $D_{coex}(g)$, $D_{cometh}(g)$, $D_{snp}(g)$, $D_{metab}(g)$, $D_{pymbms}(g)$ represent the degrees of gene g in the lignin subnetworks N^L_{coex}, N^L_{cometh}, N^L_{snp}, N^L_{metab}, and N^L_{pymbms}, respectively. The LOE *breadth* score $LOE_{breadth}(g)$ is then defined as

$$\begin{aligned} LOE_{breadth}(g) = \ & \text{bin}\left(D_{coex}(g)\right) + \text{bin}\left(D_{cometh}(g)\right) + \\ & \text{bin}\left(D_{snp}(g)\right) + \text{bin}\left(D_{metab}(g)\right) + \\ & \text{bin}\left(D_{pymbms}(g)\right) \end{aligned} \qquad (20)$$

where

$$\text{bin}(x) = \begin{cases} 1 \text{ if } x \geq 1 \\ 0 \text{ otherwise} \end{cases} \qquad (21)$$

The $LOE_{breadth}(g)$ score indicates the number of different types of lines of evidence that exist linking gene g to lignin-related genes/phenotypes.

The LOE *depth* score $LOE_{depth}(g)$ represents the total number of lines of evidence exist linking gene g to lignin-related genes/phenotypes, and is defined as

$$\begin{aligned} LOE_{depth}(g) = \ & D_{coex}(g) + D_{cometh}(g) + D_{snp}(g) + D_{metab}(g) + \\ & D_{pymbms}(g) \end{aligned} \qquad (22)$$

The GWAS LOE score $LOE_{gwas}(g)$ indicates the number of lignin-related phenotypes (metabolomic or pyMBMS) that a gene is connected to, and is defined as:

$$LOE_{gwas}(g) = D_{metab}(g) + D_{pymbms}(g) \qquad (23)$$

Distributions of the LOE scores can be seen in Figure S2. Cytoscape version 3.4.0 (Shannon et al., 2003) was used for network visualization. Expression, methylation, SNP correlation, and GWAS diagrams were created using R (R Core Team, 2017) and various R libraries (Wickham, 2007, 2009; de Vries and Ripley, 2016; Arnold, 2017; Auguie, 2017). Data parsing, wrappers and LOE score calculation was performed using Perl. Diagrams were edited to overlay certain text using Microsoft PowerPoint.

3. RESULTS AND DISCUSSION

3.1. Layered Networks, LOE Scores, and New Potential Targets

This study involved the construction of a set of networks providing different layers of information about the relationships between genes, and between genes and phenotypes, and the development of a Lines Of Evidence scoring system (LOE scores) which integrate the information in the different network layers and quantify the number of lines of evidence connecting genes to lignin-related genes/phenotypes. The GWAS network layers provide information as to which genes are potentially involved in certain functions because they contain genomic variants significantly associated with measured phenotypes. The co-methylation and co-expression networks provide information on different layers of regulatory mechanisms within the cell. The SNP correlation network provides information about possible co-evolution relationships between genes, through correlated variants across a population.

Marking known genes and phenotypes involved in lignin biosynthesis in these networks allowed for the calculation of a set of LOE (Lines Of Evidence) scores for each gene, indicating the strength of the evidence linking each gene to lignin-related functions. The breadth LOE score indicates the number of types of lines of evidence (number of layers) which connect the gene to lignin-related genes/phenotypes, whereas the depth LOE score indicates the total number of lignin-related genes/phenotypes the gene is associated with. Individual layer LOE scores (e.g., co-expression LOE score or GWAS LOE score) indicate the number of lignin-related associations the gene has within that layer.

This data layering approach differs from previous data integration methods. Mizrachi et al. (2017) integrate gene expression data with eQTN data and gene relationships from KEGG though matrix multiplication, before correlating genes' Network Based Data Integration (NBDI)-transformed values with measured traits, allowing the ranking of genes. The Mergeomics method (Shu et al., 2016) performs Marker Set Enrichment Analysis, ranking predefined sets of molecular markers based on their enrichment in a disease phenotype. Knetminer (Hassani-Pak et al., 2016; Hassani-Pak, 2017) is a web server which allows the user to search for keywords, producing lists of genes and the associations they have to annotations, genes, phenotypes, publications etc. which match the keywords and that are available in public databases. Knetminer can also produce a network view of the results. While Knetminer is also an approach which utilizes multiple lines of evidence, the main approach and the scoring systems differ. LOE requires input lists of genes and phenotypes of interest to the user, Knetminer uses gene lists and keyword searching. In terms of lines of evidence, Knetminer counts the number of "concepts" (nodes, including publications, phenotypes, annotations etc.) a gene has linking it to a keyword (Hassani-Pak et al., 2016; Hassani-Pak, 2017). However, LOE scores (particularly, breadth LOE scores) count the number of types of relationships (e.g., GWAS association, co-expression, co-methylation, variant correlation *edges*) connecting a gene to specific input genes and phenotypes related to the user's function of interest. This is thus a valuable approach to

identify new target genes based on the *relationships* of a gene to target genes/phenotypes of interest in custom-made association network layers where publically available data is not available.
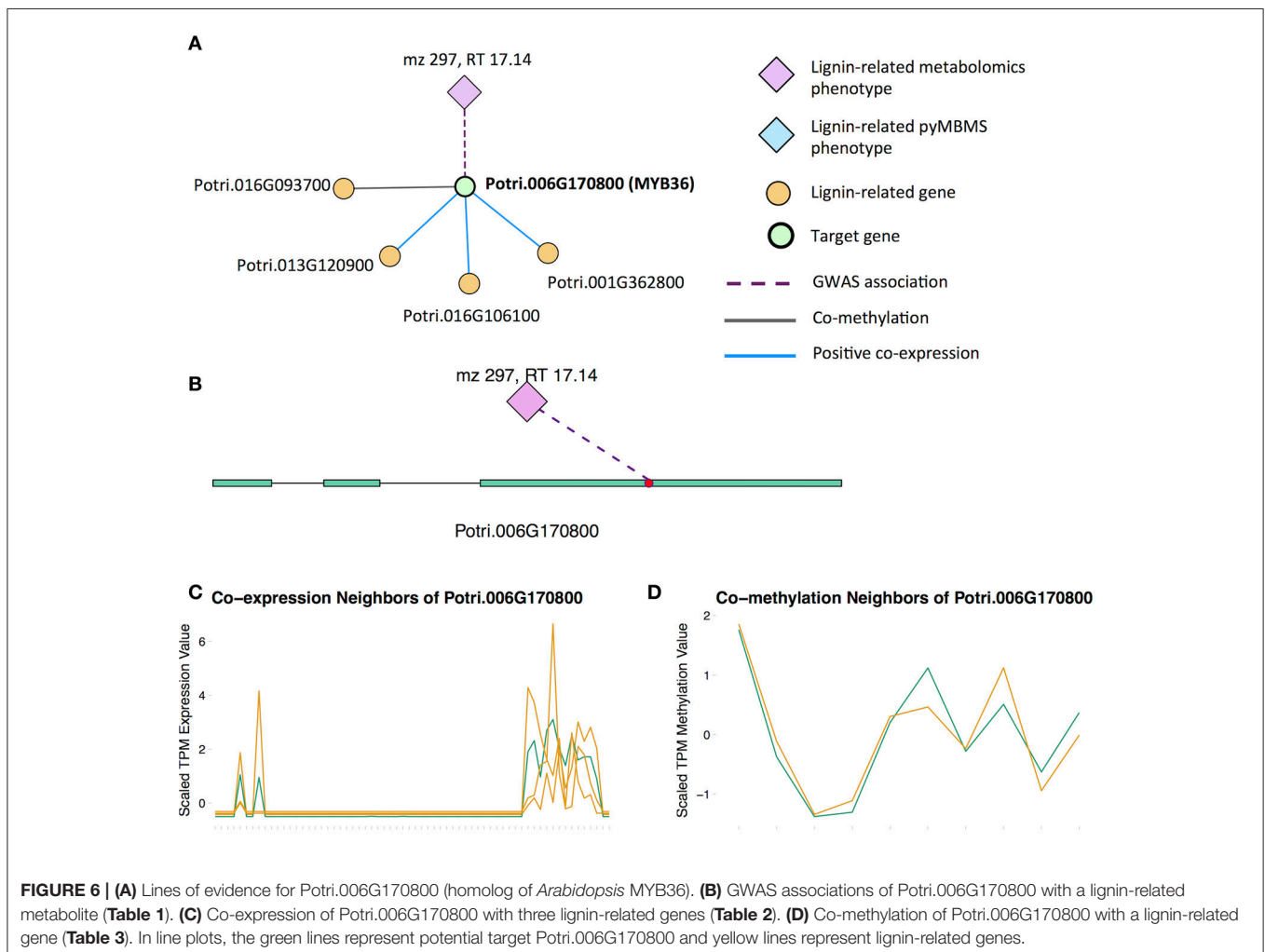
To select the top set of potential new candidate genes involved in lignin biosynthesis, genes which showed a number of different lines of evidence connecting them to lignin-related functions were identified by selecting genes with a LOE breadth score ≥3. Since the GWAS networks provide the highest resolution, most direct connections to lignin-related functions, it was also required that our potential new targets had a GWAS score ≥1. This provides a set of 375 new candidate genes potentially involved in lignin biosynthesis, identified through multiple lines of evidence (Table S4). This set of Potential New Target genes will be referred to as set of PNTs. A selection of these potential new candidates below and their annotations, derived from their *Arabidopsis* best hits, will be discussed below.

3.2. Agamous-Like Genes

Genes in the AGAMOUS-LIKE gene family are MADS-box transcription factors, many of which which have been found to play important roles in floral development (Lee et al., 2000;

Yu et al., 2002, 2004, 2017; Yoo et al., 2006; Fernandez et al., 2014). Three potential AGAMOUS-LIKE (AGL) genes are found in the set of PNTs, in particular, a homolog of *Arabidopsis* AGL8 (AT5G60910, also known as FRUITFUL), a homolog of *Arabidopsis* AGL12 (AT1G71692), and a homolog of *Arabidopsis* AGL24 (AT4G24540) and AGL22 (AT2G22540).

The first potential AGL gene in our set of PNTs is Potri.012G062300, with a breadth score of 3 and a GWAS score of 2 (**Figure 3A**), whose best *Arabidopsis thaliana* hit is AGL8 (AT5G60910). It has GWAS associations with a lignin-related metabolite (quinic acid) and a lignin pyMBMS peak (syringol) (**Figure 3C**, **Table 1**) and is co-methylated with three lignin-related genes (**Figure 3B**, **Table 3**). There is thus strong evidence for the involvement of *P. trichocarpa* AGL8 in the regulation of lignin-related functions. There is literature evidence that supports the hypothesis of AGL8's involvement in the regulation of lignin biosynthesis. A patent exists for the use of AGL8 expression in reducing the lignin content of plants (Yanofsky et al., 2004). The role of AGL8 (FUL) was described in Ferrándiz et al. (2000), in which they investigated the differences in lignin deposition in transgenic plants in which AGL8 is

FIGURE 6 | (A) Lines of evidence for Potri.006G170800 (homolog of *Arabidopsis* MYB36). **(B)** GWAS associations of Potri.006G170800 with a lignin-related metabolite (**Table 1**). **(C)** Co-expression of Potri.006G170800 with three lignin-related genes (**Table 2**). **(D)** Co-methylation of Potri.006G170800 with a lignin-related gene (**Table 3**). In line plots, the green lines represent potential target Potri.006G170800 and yellow lines represent lignin-related genes.

constitutively expressed, loss-of-function AGL8 mutants and wild-type *Arabidopsis* plants (Ferrándiz et al., 2000). In wild-type plants, a single layer of valve cells were lignified. In loss-of-function AGL8 mutants, all valve mesophyl cell layers were lignified, while in the transgenic plants, constitutive expression of AGL8 resulted in loss of lignified cells (Ferrándiz et al., 2000). This study thus showed the involvement of AGL8 in fruit lignification during fruit development.

There is evidence of other AGAMOUS-LIKE genes affecting lignin content. A study by Giménez et al. (2010) investigated TALG1, an AGAMOUS-LIKE gene in tomato, and found that TAGL1 RNAi-silenced fruits showed increased lignin content, and increased expression levels of lignin biosynthesis genes (Giménez et al., 2010). A recent study by Cosio et al. (2017)

showed that AGL15 in *Arabidopsis* is also involved in regulating lignin-related functions, in that AGL15 binds to the promotor of peroxidase PRX17, and regulates its expression (Cosio et al., 2017). In addition, PRX17 loss of function mutants had reduced lignin content (Cosio et al., 2017).

There is thus compelling evidence that various AGAMOUS-LIKE genes are involved in regulating lignin biosynthesis/deposition in plants. Two other AGAMOUS-like genes are seen in the set of PNTs, namely a homolog of *Arabidopsis* AGL12 (Potri.013G102600) and a homolog of *Arabidopsis* AGL22/AGL24 (Potri.007G115100). Potri.013G102600 (AGL12) has GWAS associations with three lignin-related metabolites, namely hydroxyphenyl lignan glycoside, coumaroyl-tremuloidin, and 3-O-caffeoyl-quinate
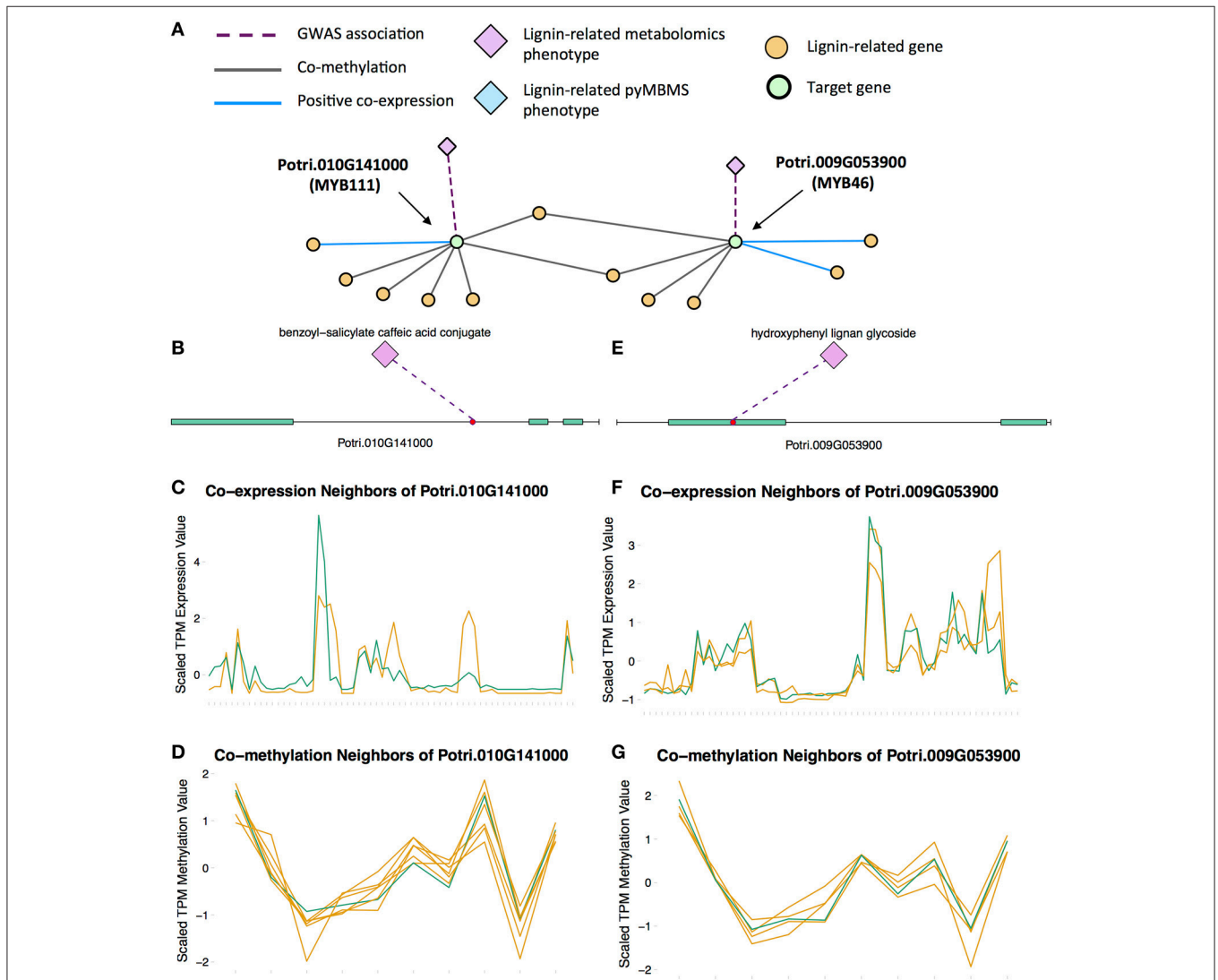
FIGURE 7 | (A) Lines of evidence for Potri.009G053900 (homolog of *Arabidopsis* MYB46) and Potri.010G141000 (homolog of *Arabidopsis* MYB111). **(B)** GWAS associations of Potri.010G141000 with a lignin-related metabolite (**Table 1**). **(C)** Co-expression of Potri.010G141000 with a lignin-related gene (**Table 2**). **(D)** Co-methylation of Potri.010G141000 with six lignin-related genes (**Table 3**). **(E)** GWAS associations of Potri.009G053900 with a lignin-related metabolite (**Table 1**). **(F)** Co-expression of Potri.009G053900 with two lignin-related genes (**Table 2**). **(G)** Co-methylation of Potri.009G053900 with four lignin-related genes (**Table 3**). In line plots, the green lines represent potential targets Potri.009G053900/Potri.010G141000 and yellow lines represent lignin-related genes.

(**Figures 4A,B**, **Table 1**). It is co-expressed with four lignin-related genes including two caffeoyl coenzyme A O-methyltransferases, a caffeate O-methyltransferase and a ferulic acid 5-hydroxylase (**Figures 4A,C**, **Table 2**) and it is co-methylated with four other lignin-related genes (**Figures 4A,D**, **Table 3**). Potri.007G115100 (AGL22/AGL24) has GWAS associations with the syringaldehyde pyMBMS phenotype and a caffeoyl conjugate metabolite (**Figures 5A,B**, **Table 1**). It also has SNP correlations with a laccase and a

nicotinamidase (**Figures 5A,C,D**, **Table 4**, Table S5). The combination of the multiple lines of multi-omic evidence thus suggest the involvement of *P. trichocarpa* homologs of *A. thaliana* AGL22/AGL24 and AGL12 in regulating lignin biosynthesis.

3.3. MYB Transcription Factors

MYB proteins contain the conserved MYB DNA-binding domain, and usually function as transcription factors.

FIGURE 8 | (A) Lines of evidence for Potri.016G078600 (homolog of *Arabidopsis* cpSRP54). **(B)** GWAS associations of Potri.016G078600 with three lignin-related metabolite (**Table 1**). **(C)** Correlations of SNPs within Potri.016G078600 with SNPs in a lignin-related gene (**Table 4**). **(D)** Co-expression of Potri.016G078600 with two lignin-related genes (**Table 2**). **(E)** Co-methylation of Potri.016G078600 with a lignin-related gene (**Table 4**). In line plots, the green lines represent potential target Potri.016G078600 and yellow lines represent lignin-related genes.

R2R3-MYBs have been found to regulate various functions, including flavonol biosynthesis, anthocyanin biosynthesis, lignin biosynthesis, cell fate, and developmental functions (Dubos et al., 2010). The set of PNTs contains several genes which are homologs of *Arabidopsis* MYB transcription factors, including homologs of *Arabidopsis* MYB66/MYB3, MYB46, MYB36, and MYB111.

There is already existing literature evidence for how some of these MYBs affect lignin biosynthesis. Liu et al. (2015) reviews the involvement of MYB transcription factors in the regulation of phenylpropanoid metabolism. MYB3 in *Arabidopsis* is known to repress phenylpropanoid biosynthesis (Zhou M. et al., 2017), and a *P. trichocarpa* homolog of MYB3 is found in our set of potential new targets. Another potential new target is the *P. trichocarpa* homolog of *Arabidopsis* MYB36 (Potri.006G170800) which is connected to lignin-related functions through multiple lines of evidence (**Figure 6**). In *Arabidopsis*, MYB36 has been found to regulate the local deposition of lignin during casparian strip formation, and *myb36* mutants exhibit incorrectly localized lignin deposition (Kamiya et al., 2015).

MYB46 is known to be a regulator of secondary cell wall formation (Zhong et al., 2007). Overexpression of MYB46 in *Arabidopsis* activates lignin, cellulose and xylan biosynthesis pathways (Zhong et al., 2007). The MYB46 homolog in *P. trichocarpa*, Potri.009G053900, is connected to lignin-related functions through multiple lines of evidence (**Figure 7A**), including a GWAS association with a hydroxyphenyl lignan glycoside (**Figure 7E**, **Table 1**), co-expression with pinoresinol reductase 1 and caffeate O-methyltransferase 1 (**Figure 7F**, **Table 2**) and co-methylation with dehydroquinate-shikimate dehydrogenase enzyme, cinnamyl alcohol dehydrogenase 9, 4-coumarate-CoA ligase activity/4CL), and caffeoyl-CoA 3-O-methyltransferase (**Figure 7G**, **Table 3**).

A MYB transcription factor in the set of PNTs which has, to our knowledge, not yet been directly associated with lignin biosynthesis is MYB111 (**Figures 7A–D**). However, with existing literature evidence, one can hypothesize that MYB111 can alter lignin content by redirecting carbon flux from flavonoids to monolignols. There is evidence that MYB111 is involved in crosstalk between lignin and flavonoid pathways. Monolignols and flavonoids are both derived from phenylalanine through the phenylpropanoid pathway (Liu et al., 2015). There is crosstalk between the signaling pathways of ultraviolet-B (UV-B) stress and biotic stress pathways (Schenke et al., 2011). In the study by Schenke et al. (2011), it was shown that under UV-B light stress, *Arabidopsis* plants produce flavonols as a UV protectant. Also, simultaniously applying the bacterial elicitor flg22, which simulates biotic stress, repressed flavonol biosynthesis genes and induced production of defense compounds including camalexin and scopoletin, as well as lignin, which provides a physical barrier preventing pathogens' entry (Schenke et al., 2011). This crosstalk involved regulation by MYB12 and MYB4 (Schenke et al., 2011). This study by Schenke et al. (2011) was performed using cell cultures. A second study (Zhou Z. et al., 2017) used *Arabidopsis* seedlings, and found that MYB111 may be involved in the crosstalk in planta (Zhou Z. et al., 2017). The multiple lines of evidence connecting the *P. trichocarpa* homolog of *Arabidopsis*

MYB111 (Potri.010G141000) to lignin related functions, in combination with the above literature evidence suggests the involvement this gene in the regulation of lignin biosynthesis by redirecting carbon flux from flavonol biosynthesis to monolignol biosynthesis, as part of the crosstalk between UV-B protection and biotic stress signaling pathways.

3.4. Chloroplast Signal Recognition Particle

Potri.016G078600, a homolog of the *Arabidopsis* chloropast signal recognition particle cpSRP54 occurs in the set of PNTs (**Figure 8**). It has a GWAS LOE score of 3, through GWAS associations with salicyl-coumaroyl-glucoside, a caffeoyl conjugate and a feruloyl conjugate (**Figure 8B**, **Table 1**, Table S4). It also has a breadth score of 4, indicating that it is linked to lignin-related genes/phenotypes though 4 different types of associations (**Figure 8**). CpSRP54 gene has been found to regulate carotenoid accumulation in *Arabidopsis* (Yu et al., 2012). CpSRP54 and cpSRP43 form a "transit complex" along with a light-harvesting chlorophyll a/b-binding protein (LHCP) family member to transport it to the thylakoid membrane (Groves et al., 2001; Schünemann, 2004). A study in *Arabidopsis* found that cpSRP43 mutants had reduced lignin content (Klenell et al., 2005). Since CpSRP54 regulates carotenoid accumulation, and cpSRP43 appears to affect lignin content, it is possible that chloroplast signal recognition particles affect lignin and carotenoid content through flux through the phenylpropanoid pathway, the common origin of both of these compounds. In fact, a gene mutation *cue1* which causes LHCP underexpression also results in reduced aromatic amino acid biosynthesis (Streatfield et al., 1999). These multiple lines of evidence, combined with the above cited literature suggests that chloroplast signal recognition particles in *P. trichocarpa* could potentially influence lignin content.

3.5. Practical Implications

The LOE method of data integration provides a useful way for biologists to identify new target genes. Any genes and phenotypes of interest that are present in the networks can be used as input to the method, and thus, the results can be tailored to the particular function of interest of the biologist. The collection of LOE scores will allow the user to rank genes in the genome based on the particular lines of evidence most appropriate to function under investigation, and in so doing, provides a shortlist of genes as targets for genetic modification (knockout/knockdown/overexpression) in order to alter the phenotype of interest. For example, AGL genes, MYB transcription factors and CpSRP genes discussed above could be seen as potential new targets for knockout/knockdown/overexpression in order to alter the lignin content of *P. trichocarpa*.

The LOE scoring method can be applied to any species for which there is multiple data types that can be represented as association networks which the scientist wishes to integrate in order to identify new candidate genes involved in a particular function. This method will be particularly useful for the analysis of new, unpublished datasets where publically available datasets/web servers would not necessarily be able to be used.

4. CONCLUDING REMARKS

This study made use of high-resolution GWAS data, combined with co-expression, co-methylation and SNP correlation networks in a multi-omic, data layering approach which has allowed the identification of new potential target genes involved in lignin biosynthesis/regulation. Various literature evidence supports the involvement of many of these new target genes in lignin biosynthesis/regulation, and these are suggested for future validation for involvement in the regulation of lignin biosynthesis. The data layering technique and LOE scoring system developed can be applied to other omic data types to assist in the generation of new hypotheses surrounding various functions of interest.

AUTHOR CONTRIBUTIONS

DW calculated methylation TPM values, constructed the networks, developed the scoring technique, performed the data layering and scoring analysis, and interpreted the results; PJ performed the outlier analysis and GWAS; MS mapped gene expression atlas reads and calculated gene expression TPM values; SD, GT, and WM lead the effort on constructing the GWAS population; TT led the leaf sample collection for GCMS-based metabolomic analyses, identified the peaks, and summarized the metabolomics data; PR did automated extraction of metabolite intensity from GCMS; MM collected the leaf samples and manually extracted the metabolite data; NZ conducted leaf sample preparation, extracted and derivatized, and analyzed the metabolites by GCMS; JS and AS generated the gene expression atlas data; SD and DM-S generated the SNP calls; RS generated the pyMBMS data; DJ conceived of and supervised the project, generated MapMan annotations, and edited the manuscript; DW, PJ, SD, DM-S, RS, TT, JS, and AS wrote the manuscript.

FUNDING

ACKNOWLEDGMENTS

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fenrg.2018.00030/full#supplementary-material

REFERENCES

Akulenko, R., and Helms, V. (2013). DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Hum. Mol. Genet.* 22, 3016–3022. doi: 10.1093/hmg/ddt158

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Arnold, J. B. (2017). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 3.4.0.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., and Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 27, 1691–1692. doi: 10.1093/bioinformatics/btr174

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.

Biswal, A. K., Hao, Z., Pattathil, S., Yang, X., Winkeler, K., Collins, C., et al. (2015). Downregulation of gaut12 in populus deltoides by rna silencing results in reduced recalcitrance, increased growth and reduced xylan and pectin in a woody biofuel feedstock. *Biotechnol. Biofuels* 8:41. doi: 10.1186/s13068-015-0218-y

Bunyavanich, S., Schadt, E. E., Himes, B. E., Lasky-Su, J., Qiu, W., Lazarus, R., et al. (2014). Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis. *BMC Med. Genomics* 7:48. doi: 10.1186/1755-8794-7-48

Busch, R., Qiu, W., Lasky-Su, J., Morrow, J., Criner, G., and DeMeo, D. (2016). Differential DNA methylation marks and gene comethylation of COPD in African-Americans with COPD exacerbations. *Respir. Res.* 17:143. doi: 10.1186/s12931-016-0459-8

Calabrese, G. M., Mesner, L. D., Stains, J. P., Tommasini, S. M., Horowitz, M. C., Rosen, C. J., et al. (2017). Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell Syst.* 4, 46–59.e4. doi: 10.1016/j.cels.2016.10.014

Climer, S., Templeton, A. R., and Zhang, W. (2014a). Allele-specific network reveals combinatorial interaction that transcends small effects in psoriasis GWAS. *PLoS Comput. Biol.* 10:e1003766. doi: 10.1371/journal.pcbi.1003766

Climer, S., Yang, W., Fuentes, L., Dávila-Román, V. G., and Gu, C. C. (2014b). A custom correlation coefficient (CCC) approach for fast identification of multi-SNP association patterns in genome-wide SNPs data. *Genet. Epidemiol.* 38, 610–621. doi: 10.1002/gepi.21833

Cosio, C., Ranocha, P., Francoz, E., Burlat, V., Zheng, Y., Perry, S. E., et al. (2017). The class III peroxidase PRX17 is a direct target of the MADS-box transcription factor AGAMOUS-LIKE15 (AGL15) and participates in lignified tissue formation. *New Phytol.* 213, 250–263. doi: 10.1111/nph.14127

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Davies, M. N., Volta, M., Pidsley, R., Lunnon, K., Dixit, A., Lovestone, S., et al. (2012). Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.* 13:R43. doi: 10.1186/gb-2012-13-6-r43

Davis, M. F., Tuskan, G. A., Payne, P., Tschaplinski, T. J., and Meilan, R. (2006). Assessment of *Populus* wood chemistry following the introduction of a Bt toxin gene. *Tree Physiol.* 26, 557–564. doi: 10.1093/treephys/26.5.557

de Vries, A., and Ripley, B. D. (2016). *ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'*. R package version 0.1-20.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., and Lepiniec, L. (2010). MYB transcription factors in *Arabidopsis*. *Trends Plant Sci.* 15, 573–581. doi: 10.1016/j.tplants.2010.06.005

Evans, L. M., Slavov, G. T., Rodgers-Melnick, E., Martin, J., Ranjan, P., Muchero, W., et al. (2014). Population genomics of *Populus trichocarpa* identifies signatures of selection and adaptive trait associations. *Nat. Genet.* 46, 1089–1096. doi: 10.1038/ng.3075

Evans, R. J., and Milne, T. A. (1987). Molecular characterization of the pyrolysis of biomass. *Energy Fuels* 1, 123–137.

Fernandez, D. E., Wang, C.-T., Zheng, Y., Adamczyk, B. J., Singhal, R., Hall, P. K., et al. (2014). The MADS-domain factors AGAMOUS-LIKE15 and AGAMOUS-LIKE18, along with SHORT VEGETATIVE PHASE and AGAMOUS-LIKE24, are necessary to block floral gene expression during the vegetative phase. *Plant Physiol.* 165, 1591–1603. doi: 10.1104/pp.114.242990

Ferrándiz, C., Liljegren, S. J., and Yanofsky, M. F. (2000). Negative regulation of the *SHATTERPROOF* genes by FRUITFULL during *Arabidopsis* fruit development. *Science* 289, 436–438. doi: 10.1126/science.289.5478.436

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344

Fridley, B. L., Lund, S., Jenkins, G. D., and Wang, L. (2012). A Bayesian integrative genomic model for pathway analysis of complex traits. *Genet. Epidemiol.* 36, 352–359. doi: 10.1002/gepi.21628

Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi: 10.1093/nar/gkw1108

Geraldes, A., Difazio, S. P., Slavov, G. T., Ranjan, P., Muchero, W., Hannemann, J., et al. (2013). A 34K SNP genotyping array for *Populus trichocarpa*: Design, application to the study of natural populations and transferability to other *Populus* species. *Mol. Ecol. Res.* 13, 306–323. doi: 10.1111/1755-0998.12056

Giménez, E., Pineda, B., Capel, J., Antón, M. T., Atarés, A., Pérez-Martín, F., et al. (2010). Functional analysis of the *Arlequin* mutant corroborates the essential role of the *Arlequin/TAGL1* gene during reproductive development of tomato. *PLoS ONE* 5:e14427. doi: 10.1371/journal.pone.0014427

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Groves, M. R., Mant, A., Kuhn, A., Koch, J., Dübel, S., Robinson, C., and Sinning, I. (2001). Functional characterization of recombinant chloroplast signal recognition particle. *J. Biol. Chem.* 276, 27778–27786. doi: 10.1074/jbc.M103470200

Hassani-Pak, K. (2017). *KnetMiner - An Integrated Data Platform for Gene Mining and Biological Knowledge Discovery*. Ph.D. thesis, Universität Bielefeld.

Hassani-Pak, K., Castellote, M., Esch, M., Hindle, M., Lysenko, A., Taubert, J., et al. (2016). Developing integrated crop knowledge networks to advance candidate gene discovery. *Appl. Trans. Genomics* 11, 18–26. doi: 10.1016/j.atg.2016.10.003

Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 15:182. doi: 10.1186/1471-2105-15-182

Kamiya, T., Borghi, M., Wang, P., Danku, J. M., Kalmbach, L., Hosmani, P. S., et al. (2015). The MYB36 transcription factor orchestrates Casparian strip formation. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10533–10538. doi: 10.1073/pnas.1507691112

Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Kim, D., Li, R., Dudek, S. M., and Ritchie, M. D. (2013). ATHENA: identifying interactions between different levels of genomic data associated with cancer clinical outcomes using grammatical evolution neural network. *Biodata Mining* 6:23. doi: 10.1186/1756-0381-6-23

Kim, D., Shin, H., Song, Y. S., and Kim, J. H. (2012). Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. Biomed. Informatics* 45, 1191–1198. doi: 10.1016/j.jbi.2012.07.008

Klenell, M., Morita, S., Tiemblo-Olmo, M., Mühlenbock, P., Karpinski, S., and Karpinska, B. (2005). Involvement of the chloroplast signal recognition particle cpSRP43 in acclimation to conditions promoting photooxidative stress in *Arabidopsis*. *Plant Cell Physiol.* 46, 118–129. doi: 10.1093/pcp/pci010

Lee, H., Suh, S.-S., Park, E., Cho, E., Ahn, J. H., Kim, S.-G., et al. (2000). The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in *Arabidopsis*. *Genes Dev.* 14, 2366–2376. doi: 10.1101/gad.813600

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* 49, 764–766. doi: 10.1016/j.jesp.2013.03.013

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, Y., Tschaplinski, T. J., Engle, N. L., Hamilton, C. Y., Rodriguez, M., Liao, J. C., et al. (2012). Combined inactivation of the *Clostridium cellulolyticum* lactate and malate dehydrogenase genes substantially increases ethanol yield from cellulose and switchgrass fermentations. *Biotechnol. Biofuels* 5:2. doi: 10.1186/1754-6834-5-2

Li, Z., and Trick, H. N. (2005). Rapid method for high-quality RNA isolation from seed endosperm containing high levels of starch. *Biotechniques* 38:872.

Liu, J., Osbourn, A., and Ma, P. (2015). MYB transcription factors as regulators of phenylpropanoid metabolism in plants. *Mol. Plant* 8, 689–708. doi: 10.1016/j.molp.2015.03.012

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., et al. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ.* 37, 1250–1258. doi: 10.1111/pce.12231

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., et al. (2018). The sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 93, 338–354. doi: 10.1111/tpj.13781

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A.,et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

McKown, A. D., Klápště, J., Guy, R. D., Geraldes, A., Porth, I., Hannemann, J., et al. (2014). Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytol.* 203, 535–553. doi: 10.1111/nph.12815

Mizrachi, E., Verbeke, L., Christie, N., Fierro, A. C., Mansfield, S. D., Davis, M. F., et al. (2017). Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. *Proc. Natl. Acad. Sci. U.S.A.* 114, 1195–1200. doi: 10.1073/pnas.1620119114

Muchero, W., Guo, J., DiFazio, S. P., Chen, J.-G., Ranjan, P., Slavov, G. T., et al. (2015). High-resolution genetic mapping of allelic variants associated with cell wall chemistry in *Populus*. *BMC Genomics* 16:24. doi: 10.1186/s12864-015-1215-z

Price-Whelan, A. M., and Foreman-Mackey, D. (2017). schwimmbad: a uniform interface to parallel processing pools in Python. *J. Open Source Softw.* 2:357. doi: 10.21105/joss.00357

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., et al. (2006). The path forward for biofuels and biomaterials. *Science*, 311, 484–489. doi: 10.1126/science.1114736

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 16, 85–97. doi: 10.1038/nrg3868

Sannigrahi, P., Ragauskas, A. J., and Tuskan, G. A. (2010). Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels Bioprod. Biorefin.* 4, 209–226. doi: 10.1002/bbb.206

Schenke, D., Böettcher, C., and Scheel, D. (2011). Crosstalk between abiotic ultraviolet-B stress and biotic (flg22) stress signalling in *Arabidopsis* prevents flavonol accumulation in favor of pathogen defence compound production. *Plant Cell Environ.* 34, 1849–1864. doi: 10.1111/j.1365-3040.2011.02381.x

Schünemann, D. (2004). Structure and function of the chloroplast signal recognition particle. *Curr. Genet.* 44, 295–304. doi: 10.1007/s00294-003-0450-z

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shu, L., Zhao, Y., Kurt, Z., Byars, S. G., Tukiainen, T., Kettunen, J., et al. (2016). Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics* 17:874. doi: 10.1186/s12864-016-3198-9

Slavov, G. T., DiFazio, S. P., Martin, J., Schackwitz, W., Muchero, W., Rodgers-Melnick, E., et al. (2012). Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol.* 196, 713–725. doi: 10.1111/j.1469-8137.2012.04258.x

Streatfield, S. J., Weber, A., Kinsman, E. A., Häusler, R. E., Li, J., Post-Beittenmiller, D., et al. (1999). The phosphoenolpyruvate/phosphate translocator is required for phenolic metabolism, palisade cell development, and plastid-dependent nuclear gene expression. *Plant Cell* 11, 1609–1621.

Sykes, R., Yung, M., Novaes, E., Kirst, M., Peter, G., and Davis, M. (2009). High-throughput screening of plant cell-wall composition using pyrolysis molecular beam mass spectroscopy. *Biofuels Methods Protoc.* 581, 169–183. doi: 10.1007/978-1-60761-214-8_12

Timm, C. M., Pelletier, D. A., Jawdy, S. S., Gunter, L. E., Henning, J. A., Engle, N., et al. (2016). Two poplar-associated bacterial isolates induce additive favorable responses in a constructed plant-microbiome system. *Front. Plant Sci.* 7:497. doi: 10.3389/fpls.2016.00497

Tschaplinski, T. J., Standaert, R. F., Engle, N. L., Martin, M. Z., Sangha, A. K., Parks, J. M., et al. (2012). Down-regulation of the caffeic acid *O*-methyltransferase gene in switchgrass reveals a novel monolignol analog. *Biotechnol. Biofuels* 5:71. doi: 10.1186/1754-6834-5-71

Tuskan, G., Slavov, G., DiFazio, S., Muchero, W., Pryia, R., Schackwitz, W., et al. (2011). *Populus* resequencing: towards genome-wide association studies. *BMC Proc.* 5:I21. doi: 10.1186/1753-6561-5-S7-I21

Tuskan, G., West, D., Bradshaw, H. D., Neale, D., Sewell, M., Wheeler, N., et al. (1999). Two high-throughput techniques for determining wood properties as part of a molecular genetics analysis of hybrid poplar and loblolly pine. *Appl. Biochem. Biotechnol.* 77, 55–65. doi: 10.1385/ABAB:77:1-3:55

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., et al. (2013). From fastQ data to high-confidence variant Calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* 43, 11.10.1-33. doi: 10.1002/0471250953.bi1110s43

van Dongen, S. (2008). Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.* 30, 121–141. doi: 10.1137/040608635

van Dongen, S. (2000). *Graph Clustering by Flow Simulation.* Ph.D. thesis, University of Utrecht, May 2000.

Vanholme, R., Demedts, B., Morreel, K., Ralph, J., and Boerjan, W. (2010). Lignin biosynthesis and structure. *Plant Physiol.* 153, 895–905. doi: 10.1104/pp.110.155119

Vining, K. J., Pomraning, K. R., Wilhelm, L. J., Priest, H. D., Pellegrini, M., Mockler, T. C., et al. (2012). Dynamic DNA cytosine methylation in the *Populus trichocarpa* genome: tissue-level variation and relationship to gene expression. *BMC Genomics* 13:27. doi: 10.1186/1471-2164-13-27

Wagner, G. P., Kin, K., and Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* 131, 281–285. doi: 10.1007/s12064-012-0162-3

Wickham, H. (2007). Reshaping data with the reshape package. *J. Stat. Softw.* 21, 1–20. doi: 10.18637/jss.v021.i12

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York, NY: Springer-Verlag.

Yanofsky, M. F., Liljegren, S., and Ferrandiz, C. (2004). *Selective control of lignin biosynthesis in transgenic plants.* US Patent 6,768,042. Available online at: https://patents.google.com/patent/US6768042B2/en

Yoo, S. K., Lee, J. S., and Ahn, J. H. (2006). Overexpression of *AGAMOUS-LIKE 28* (*AGL28*) promotes flowering by upregulating expression of floral promoters within the autonomous pathway. *Biochem. Biophys. Res. Commun.* 348, 929–936. doi: 10.1016/j.bbrc.2006.07.121

Yu, B., Gruber, M. Y., Khachatourians, G. G., Zhou, R., Epp, D. J., Hegedus, D. D., et al. (2012). Arabidopsis cpSRP54 regulates carotenoid accumulation in *Arabidopsis* and *Brassica napus*. *J. Exp. Bot.* 63, 5189–5202. doi: 10.1093/jxb/ers179

Yu, H., Ito, T., Wellmer, F., and Meyerowitz, E. M. (2004). Repression of AGAMOUS-LIKE 24 is a crucial step in promoting flower development. *Nat. Genet.* 36, 157–161. doi: 10.1038/ng1286

Yu, H., Xu, Y., Tan, E. L., and Kumar, P. P. (2002). AGAMOUS-LIKE 24, a dosage-dependent mediator of the flowering signals. *Proc. Natl. Acad. Sci. U.S.A.* 99, 16336–16341. doi: 10.1073/pnas.212624599

Yu, X., Chen, G., Guo, X., Lu, Y., Zhang, J., Hu, J., et al. (2017). Silencing *SlAGL6*, a tomato *AGAMOUS-LIKE6* lineage gene, generates fused sepal and green petal. *Plant Cell Rep.* 36, 959–969. doi: 10.1007/s00299-017-2129-9

Zhao, Q., Zeng, Y., Yin, Y., Pu, Y., Jackson, L. A., Engle, N. L., et al. (2015). Pinoresinol reductase 1 impacts lignin distribution during secondary

cell wall biosynthesis in Arabidopsis. *Phytochemistry* 112, 170–178. doi: 10.1016/j.phytochem.2014.07.008

Zhong, R., Richardson, E. A., and Ye, Z.-H. (2007). The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in *Arabidopsis. Plant Cell* 19, 2776–2792. doi: 10.1105/tpc.107.0 53678

Zhou, M., Zhang, K., Sun, Z., Yan, M., Chen, C., Zhang, X., et al. (2017). LNK1 and LNK2 corepressors interact with the MYB3 transcription factor in phenylpropanoid biosynthesis. *Plant Physiol.* 174, 1348–1358.

Zhou, Z., Schenke, D., Miao, Y., and Cai, D. (2017). Investigation of the crosstalk between the flg22 and the UV-B-induced flavonol pathway in *Arabidopsis thaliana* seedlings. *Plant Cell Environ.* 40, 453–458. doi: 10.1111/pce.12869