New Phytologist

Research

Systems genetic analysis of lignin biosynthesis in *Populus tremula*

Mikko Luomaranta[1], Carolin Grones[1], Shruti Choudhary[2], Ana Milhinhos[1], Teitur Ahlgren Kalman[1], Ove Nilsson[2], Kathryn M. Robinson[1], Nathaniel R. Street[1,3] and Hannele Tuominen[2]

[1]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, 90187, Umeå, Sweden; [2]Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre, Swedish University of Agricultural Sciences, 90183, Umeå, Sweden; [3]SciLifeLab, Umeå University, 90187, Umeå, Sweden

Author for correspondence:
Hannele Tuominen
Email: hannele.tuominen@slu.se

Received: *28 January 2024*
Accepted: *2 July 2024*

New Phytologist (2024) **243:** 2157–2174
doi: 10.1111/nph.19993

Key words: aspen, eQTL, GWAS, HD-Zip III, lignin biosynthesis, *Populus*, wood formation.

Summary

- The genetic control underlying natural variation in lignin content and composition in trees is not fully understood. We performed a systems genetic analysis to uncover the genetic regulation of lignin biosynthesis in a natural 'SwAsp' population of aspen (*Populus tremula*) trees.
- We analyzed gene expression by RNA sequencing (RNA-seq) in differentiating xylem tissues, and lignin content and composition using Pyrolysis-GC-MS in mature wood of 268 trees from 99 genotypes.
- Abundant variation was observed for lignin content and composition, and genome-wide association study identified proteins in the pentose phosphate pathway and arabinogalactan protein glycosylation among the top-ranked genes that are associated with these traits. Variation in gene expression and the associated genetic polymorphism was revealed through the identification of 312 705 local and 292 003 distant expression quantitative trait loci (eQTL). A co-expression network analysis suggested modularization of lignin biosynthesis and novel functions for the lignin-biosynthetic CINNAMYL ALCOHOL DEHYDROGENASE 2 and CAFFEOYL-CoA O-METHYLTRANSFERASE 3. *PHENYLALANINE AMMONIA LYASE 3* was co-expressed with *HOMEOBOX PROTEIN 5* (HB5), and the role of HB5 in stimulating lignification was demonstrated in transgenic trees.
- The systems genetic approach allowed linking natural variation in lignin biosynthesis to trees´ responses to external cues such as mechanical stimulus and nutrient availability.

Introduction

Lignin is a phenolic polymer that is synthesized not only in vascular plants as part of normal plant development but also in response to various external stimuli (Chantreau & Tuominen, 2022). In xylem tissues, lignin consists of three main types of monolignol subunits, the guaiacyl (G), *p*-hydroxyphenyl (H), and syringyl (S) lignin, which are deposited through oxidative coupling reactions into the secondary cell walls (Boerjan *et al.*, 2003; Vanholme *et al.*, 2019). The chemically robust structure of the lignin polymer is a major contributor to the recalcitrance of woody biomass during the bioprocessing of forest tree feedstocks (Ragauskas *et al.*, 2014; Meng *et al.*, 2017; De Meester *et al.*, 2022). Good knowledge on the monolignol biosynthetic pathway and its transcriptional regulation (Zhang *et al.*, 2020) has during the last 15 yr given rise to extensive efforts to improve feedstock processability by genetic engineering, but which were in most cases hampered by a trade-off with biomass production (Chanoca *et al.*, 2019; De Meester *et al.*, 2022). More subtle approaches are therefore needed to improve wood processability. Exploiting natural variation in lignin biosynthesis is an alternative approach to provide feedstocks that are better suited for bioprocessing. Combined with genomic tools, such as genome-wide association studies (GWAS) and gene expression analysis, natural variation can also reveal novel factors underlying lignin traits.

In various *Populus* species, considerable variation has been reported in the lignin content of natural variants (Studer *et al.*, 2011; Fahrenkrog *et al.*, 2017; Furches *et al.*, 2019; Escamez *et al.*, 2023). This variation has enabled the identification of associations between lignin traits and the underlying genetic variants by GWAS. Both Guerra *et al.* (2013) and Fahrenkrog *et al.* (2017) detected associations for the S : G ratio in the gene region of *FERULATE-5-HYDROXYLASE 3*. The 5-enolpyruvylshikimate 3-phosphate synthase gene (EPSP) was identified by Xie *et al.* (2018) as a novel regulator of lignin content using a linkage disequilibrium-based association mapping. EPSP was shown to control the accumulation of phenylpropanoid pathway metabolites, possibly through its function as a transcriptional repressor of a SLEEPER-like gene which itself is a repressor of the *Populus* homolog for the Arabidopsis secondary cell wall master regulator MYB46 (Xie *et al.*, 2018). More recent GWAS studies in *Populus* revealed large numbers of genes associated with various lignin traits (Bryant *et al.*, 2023; Li

et al., 2023). However, the identified variants collectively explain only a small part of the variation in lignin content or composition, suggesting a complex genetic architecture with many genes having a small effect on lignification, and thus advocating for further GWAS in various populations and species.

Variation in global gene expression provides a further understanding of the genetic architecture of complex traits. This can be achieved through expression quantitative trait locus (eQTL) analysis (Nica & Dermitzakis, 2013), which utilizes the same principle as GWAS but where associations are detected between gene expression and genetic markers such as single-nucleotide polymorphisms (SNPs). Further exploration of the transcriptome often includes the construction of gene co-expression networks (Stuart *et al.*, 2003), which can facilitate the identification of regulatory relationships and the identification of biological function through the use of enrichment tests of, for example, Gene Ontology (GO) in clusters of co-expressed sets of genes. This approach has been used successfully to study lignin traits in various species. An eQTL analysis, combined with functional studies, identified MYB125 as a novel transcription factor regulating the expression of lignin-biosynthetic genes and lignin content in *Populus deltoides* (Balmant *et al.*, 2020). In another *eQTL analysis of P. deltoides* × *simonii* parents and hybrids, a calmodulin protein *PdCaM247* not only correlated in expression but also interacted with *MYB156*, *a* negative regulator of lignin biosynthesis (Zhang *et al.*, 2023). An eQTL analysis with a focus on the *HYDROXYCINNAMOYLTRANSFERASE* (*HCT*) gene family supported the association of *HCT2* with the accumulation of the phenylpropanoid pathway metabolites *cis*- and *trans*-3-*O*-caffeoylquinic acids in *Populus trichocarpa* (Zhang *et al.*, 2018). These studies have pioneered the use of systems genetics to identify new factors influencing lignin biosynthesis, encouraging further integration of the different genetic methods. They have also highlighted that the loci explaining natural variation in lignin vary between species and populations.

In this study, we performed a systems genetic analysis, integrating eQTL, gene co-expression network, and GWAS analyses to reveal novel regulatory aspects of lignification in a collection of aspen (*Populus tremula*) trees sampled across the distribution range in Sweden and grown in a common garden. The lignin content and composition, analyzed by Pyrolysis-GC-MS (Py-GC-MS), as well as gene expression varied significantly within the collection. Co-expression network analyses revealed the regulatory network of lignin biosynthesis, validated by transgenic modification of the HD-ZIP III family member *HOMEOBOX PROTEIN 5* (*HB5*). Furthermore, integration of the GWAS and eQTL analyses revealed putative novel loci controlling the accumulation of G- and S-type lignin and their association with transcriptional regulation.

Materials and Methods

Tree growth and collection of the material

This study utilized aspen (*Populus tremula* L.) trees from the Swedish Aspen (SwAsp) collection (Luquez *et al.*, 2008). The collection consists of 112 genetically unrelated genotypes, originating from 12 locations across Sweden, which show no population structure (Wang *et al.*, 2018). The trees were grown in randomized blocks in a common garden located in Ekebo (13.1°E, 55.9°N). The current study assayed shoots sprouting from the root stool of trees that were cut down in the winter of 2013–14 (Escamez *et al.*, 2023). Five-year-old shoots, representing 99 SwAsp genotypes, were sampled between 1 July and 3 July 2019, between 09:00 h and 15:00 h. The shoots ranged from 10 to 86 mm in diameter. Stem disks of 3–5 cm thickness were collected *c.* 20 cm above the base of the stem, transported on dry ice, and stored at −80°C.

Selected GWAS results were validated in the Umeå Aspen (UmAsp) collection (Fracheboud *et al.*, 2009), which consists of 227 *P. tremula* genotypes originating from the surroundings of Umeå, Sweden. Clonal replicates of the collection were grown for 10–13 years in a randomized block design in a common garden in Sävar (20.6°E, 63.9°N). Micro-cores were taken with the Trephor tool (Rossi *et al.*, 2006) from the base of the stems 10–20 cm above the ground on 23 May 2022. The cores were oven-dried at 60°C for 72 h, and processed and analyzed for wood chemistry using Pyrolysis-GC-MS.

Analysis of wood chemical composition with Pyrolysis-GC-MS (Py-GC-MS)

A quarter segment from the SwAsp stem disks and the micro-cores from the UmAsp trees were debarked, and the wood was cut into small pieces and freeze-dried. Dried wood pieces were homogenized by coarse milling (Retsch ZM200 centrifugal mill, Retsch GmbH, Germany) and sieved (Retsch AS200) into a particle size of 0.1 mm. Sixty- to seventy micrograms of homogenized wood powder was loaded into an autosampler (PY-2020iD and AS-1020E; Frontier Labs, Japan), and a sub-sample (*c.* 1 µg) was fed into the pyrolizer (Agilent, 7890A/5975C; Agilent Technologies AB, Sundbyberg, Sweden). Following pyrolysis, the samples were separated along a DB-5MS capillary column (30 m × 0.25 mm i.d., 0.25-µm-film thickness; J&W, Agilent Technologies) and scanned by a mass spectrometer along the *m/z* range 35–250. The number of replicates is given in Supporting Information Table S1. The Py-GC-MS reveals the relative contents of carbohydrates, lignin subunits (S, G, H), other phenolic compounds (P; generic benzene derivatives without OH group on the aromatic ring, most probably originated from lignin), and unknown compounds as a percentage of the sum of the area of the peaks corresponding to their respective pyrolytic products in relation to the total GC peak area (Gerber *et al.*, 2012). The relative total lignin content was calculated by summing up S, G, H, and P contents. Py-GC-MS provides relative abundances for the secondary cell wall components in a manner that is comparable to NMR (Renström *et al.*, 2024).

RNA extraction

Another quarter segment from the SwAsp stem disks was used for RNA extraction. Frozen material was debarked, and the differentiating xylem was collected by scraping from the exposed wood

surface to the depth of 2–4 mm. The scraped material was ground in liquid nitrogen using a mortar and pestle. RNA was extracted with the Trizol-chloroform method (Chomczynski & Sacchi, 2006), followed by treatment with DNase (Ambion DNase I, Thermo Fisher Scientific, Stockholm, Sweden), purified with MinElute (Qiagen) followed with additional purification steps using RNA precipitation solution (1.2 M NaCl; 0.8 M Trisodium Citrate) and isopropanol (1 : 1). The sample purity and quantity were assessed using Qubit and NanoDrop (Thermo Fisher Scientific). RNA integrity was determined using Bioanalyzer (Agilent).

RNA sequencing and data preprocessing

RNA sequencing (RNA-seq) was performed by Novogene using an Illumina NovaSeq 6000 sequencer. A target 12M 150-bp paired-end reads were produced per sample. The RNA-seq data were preprocessed following an in-house pipeline (Delhomme *et al.*, 2014). Unless specified, default parameters were used for all tools. Briefly, sequence quality was assessed using MultiQC (v.1.8.0; Ewels *et al.*, 2016), ribosomal RNA reads were filtered using sortMeRNA (v.2.1b; Kopylova *et al.*, 2012), adapters, low-quality bases and reads were removed using Trimmomatic (v.0.39; Bolger *et al.*, 2014). After filtering and rRNA removal, the quality was assessed again using MultiQC (v.1.8.0; Ewels *et al.*, 2016). Reads were then aligned against the *Populus tremula* v.2.2 genome assembly (Robinson *et al.*, 2024) using Star (v.2.7.2; Dobin *et al.*, 2013). Transcript abundance was quantified using Salmon (v.0.14.1; Patro *et al.*, 2017).

eQTL analysis

The RNA-seq data were used for eQTL analysis in the SwAsp collection. The sequenced collection consists of 99 unrelated genotypes with 6806 717 bi-allelic SNPs (Robinson *et al.*, 2024). Similar to Mähler *et al.* (2017), SNPs located further than 2 kbp from the associated gene were considered intergenic, and SNPs located within the region of a gene (2-kbp flanking, 5′/3′ UTR, exon, intron) were considered to be associated with that gene.

Gene expression data from RNA-seq was prefiltered to remove any genes with no expression, followed by transformation of the expression data using the VST function implemented in DESeq2 (v.1.42; Love *et al.*, 2014). Mean normalized gene expression values per genotype were used for the association mapping. Expression QTL (eQTL) mapping was performed using the R package Matrix eQTL (v.4.0.0; Shabalin, 2012) according to Mähler *et al.* (2017). The eQTLs were categorized as local if the SNPs were within 1 Mbp of the associated gene and distant if the distance exceeded 1 Mbp.

Gene co-expression network

Thirteen network inference methods were computed using the Seidr crowd network tool (v.0.14.4; Schiffthaler *et al.*, 2023). The network inference included use of the following implemented algorithms: ARACNE (Margolin *et al.*, 2006), CLR (Faith *et al.*, 2007), Elastic Net ensembles (ELNET) (Ruyssinck *et al.*, 2014), GENIE3 (Huynh-Thu *et al.*, 2010), linear SVM (LLR) (Ruyssinck *et al.*, 2014), mutual information (MI) (Song *et al.*, 2012), Narromi (Zhang *et al.*, 2013), partial correlation PCor, Pearson correlation, PLSNET (Guo *et al.*, 2016), Spearman, TIGRESS (Haury *et al.*, 2012), and TOM similarity (Langfelder & Horvath, 2008). For each symmetric edge pair, the one with the highest score was kept in case of nonsymmetrical scoring by the algorithm. Networks were aggregated according to the inverse rank product method (Zhong *et al.*, 2014), and the edges were filtered according to a noise-corrected backbone at a sigma of 2.32 (Coscia & Neffke, 2017). Network clustering was performed using InfoMap (v.1.8.0; Rosvall & Bergstrom, 2008) using a Markov time of 1 with default settings. Node centrality statistics were calculated using Seidr.

A lignin subnetwork was inferred from the whole-transcriptome co-expression network by selecting the lignin-biosynthetic genes and their first-degree neighbors. *P. tremula* lignin-biosynthetic genes were identified on the basis of the highest Blastn *e*-value to the *P. trichocarpa* lignin-biosynthetic genes listed in Shi *et al.* (2010). In addition, correlation values (exceeding Spearman correlation $R > 0.3$) between the Py-GC-MS traits and gene expression were included as edges in the lignin subnetwork. Transcription factors within the network were identified using the annotations at PlantGenIE.org (Sundell *et al.*, 2015).

Genome-wide association studies

The GWAS pipeline included a set of quality control steps and the estimation of a best linear unbiased predictor (BLUP) phenotypic value for each SwAsp genotype and for each trait in the GWAS. Briefly, phenotypic outliers were identified and non-normally distributed random effects and error terms were Ordered Quantile transformed with bestNormalize package in R (v.1.6.1; Peterson & Cavanaugh, 2019). Phenotypic BLUP values were calculated according to Wang *et al.* (2018). The genotype was considered as random effect, and the field block was considered as fixed effect. The same set of 6806 717 bi-allelic SNPs (Robinson *et al.*, 2024) was used as for the eQTL analysis.

The GWAS was conducted using a linear mixed model in Gemma (v.0.98.1; Zhou & Stephens, 2012). The GWAS model included the use of two covariates: latitude of origin, which was included to remove any association that could have been caused by the origin of the genotypes (Luquez *et al.*, 2008), and relatedness matrix of all individuals, as described previously (Mähler *et al.*, 2020). GWAS results were plotted as Manhattan plots generated using the R package qqman (Turner, 2014). Boxplots displaying the allelic effect on a trait were generated using the R package ggplot2 (Wickham, 2011). Figures visualizing the genomic regions for the SNPs identified in the GWAS were created using custom build of JBrowse (Buels *et al.*, 2016).

False discovery rate (FDR) for associations was calculated using q-value package in R (v.2.22.0; Storey *et al.*, 2020) according to the Benjamini–Hochberg procedure (Storey & Tibshirani, 2003). The phenotypic variance explained (PVE%) was calculated according to Wang *et al.* (2018).

The GWAS results were validated for two selected alleles in the UmAsp population (Fracheboud *et al.*, 2009). The re-sequencing data and SNP calling of UmAsp are presented in Robinson *et al.* (2024). Of the 229 UmAsp trees, 59 were selected based on the associated alleles for the most significant candidate genes for G-type (Potra2n11c22479) and S-type lignin (Potra2n1c3762) from the GWAS. For the SNP (chr11_1139060_C_G) associating with G-type lignin, 37 genotypes were homozygous for the major allele, 21 heterozygous, and 1 homozygous for the minor allele. For the SNP (chr1_48493303_A_G) associating with S-type lignin, 26 were homozygous for the major allele, 8 heterozygous, and 25 homozygous for the minor allele.

Genetic analyses

The repeatability of the assumed upper bound estimate of broad-sense heritability was calculated for gene expression and phenotype data (Dohm, 2002). Estimates of clonal heritability were calculated using the repeatability function of the R package HERITABILITY (v.1.3.0; Kruijer *et al.*, 2015). The analysis was restricted to clones with at least three replicates.

To estimate population differentiation, Q_{ST} was calculated according to the formula:

$$Q_{ST} = \frac{V_{between}}{V_{between} + 2V_{within}}$$

where $V_{between}$ describes the variance between the populations and V_{within} the residual genetic variance among genotypes within a population. Calculations were performed using the lmer function from the R package LME4 (v.1.1-33; Bates *et al.*, 2015) according to Mähler *et al.* (2017).

Statistical analyses

Correlations between the expression of the lignin-biosynthetic genes were calculated using the base R function 'cor.' Descriptive statistics, including mean, SD, and coefficient of variation for wood chemical composition were calculated with PRISM GRAPHPAD 9 software (v.9.5.0 for Windows). Heatmaps were created using the R package PHEATMAP (Kolde, 2015).

For phylogenetic trees, amino acid sequences were retrieved from PlantGenIE.org and aligned using Multiple Sequence Comparison by MUltiple Sequence Comparison by Log-Expectation (MUSCLE) (Edgar, 2004). The trees were inferred using a Maximum likelihood tree using the Bootstrap method with 1000 Bootstrap replications using the Jones–Taylor–Thornton substitution model (Jones *et al.*, 1992). Scales were drawn with branch lengths measured in the number of substitution sites.

JASPAR motif analysis

Transcription factor motifs were identified in the promoters of lignin-biosynthetic genes. First, transcription factor binding motifs from the JASPAR 2020 CORE_plant database (Fornes *et al.*, 2020) were mapped to the *P. tremula* reference genome (Robinson *et al.*, 2024) using FIMO (Grant *et al.*, 2011) from the MEME SUITE (v.5.0.5; Bailey *et al.*, 2015). A promoter region was defined as the sequence 2000-bp upstream from the transcription start site. The intersection between mapped motifs and the promoter regions was identified using BEDTOOLS (v.2.29.0; Quinlan & Hall, 2010). The promoters were compared for their presence or absence of mapped motifs, and clustered based on their average Jaccard distance using RVENN (v.1.1.0; Akyol, 2019). For the *PHENYLALANINE AMMONIA LYASE 3* (*PAL3*) promoter analysis, JASPAR 2014 CORE plant database was used (Rauluseviciute *et al.*, 2024).

Generation, wood chemistry, and qRT-PCR of RNAi lines

To generate the RNAi constructs for hybrid aspen (*Populus tremula* x *P. tremuloides*) *HB5* (Potrx046843g13917) and *HB6* (Potrx065083g25217), a 274-bp fragment was amplified from hybrid aspen cDNA using forward (agaaagctgggtGAAGGTGG TGGTTCAATTATAC) and reverse primers (aaaaagcaggctC-CATCCCTCATCACTAAATCC). The resulting fragment was recombined through pDONOR201 to pK7GWIWG2(I) which was transformed into hybrid aspen clone T89 according to Nilsson *et al.* (1992). Two transgenic lines (5AA and 5AB) were grown, along with wild-type (WT) (T89), in the glasshouse under long-day conditions (18 h : 6 h, light : dark and relative humidity of 50–70%) and fertilized once a week with Rika-S (Weibulls, Åby, Sweden). After 2 months of growth, stem height and diameter at the base were measured, and material was collected for quantitative reverse transcription polymerase chain reaction and Py-GC-MS analysis.

Stem pieces, *c.* 10 cm long, of 4–5 biological replicates, were harvested, immediately frozen, and stored at −80°C until further use. After peeling off the bark, differentiating xylem tissue was scraped and homogenized to a fine powder with liquid nitrogen. Around 100 mg of tissue was taken for the total RNA extraction using the Spectrum™ Plant Total RNA Kit (Cat# STRN250; Sigma-Aldrich) and On-Column DNase I Digest Set (Cat# DNASE70; Sigma-Aldrich). Nucleotide purity and concentration were verified using NanoDrop™ Spectrophotometer (Thermo Scientific). For cDNA synthesis, 1 µg of total RNA for each sample was utilized for a reverse transcription reaction using the iScript™ cDNA Synthesis Kit (Bio-Rad). Primers for quantitative reverse transcription polymerase chain reaction were designed using the PRIMER3 web server (v.4.1.0; Untergasser *et al.*, 2012). The primers for *HB5* (Potrx046843g13917) were 5′- TGGTTTTCGCATCATTCCCC-3′ and 5′-AGAAGCGA GGTCCAAGGTAC-3′ and for *HB6* (Potrx065083g25217) 5′-GCATCCAGTGATCATTCTGCT-3′ and 5′-ATGCCACCC TCTGAACTGAA-3′. Ubiquitin (Potra2n1c3635) was used as the reference gene using primers 5′-AGATGTGCTGTTCA TGTTGTCC-3′ and 5′-ACAGCCACTCCAAACAGTACC-3′. The polymerase chain reaction comprised 1.5 µl of 5× diluted cDNA, 1× LightCycler 480® SYBR® Green I Master (Roche), 100–200 nmol each of forward, and reverse primer in a total volume of 15 µl. The polymerase chain reaction program included an initial denaturation at 95°C for 3 min, followed by

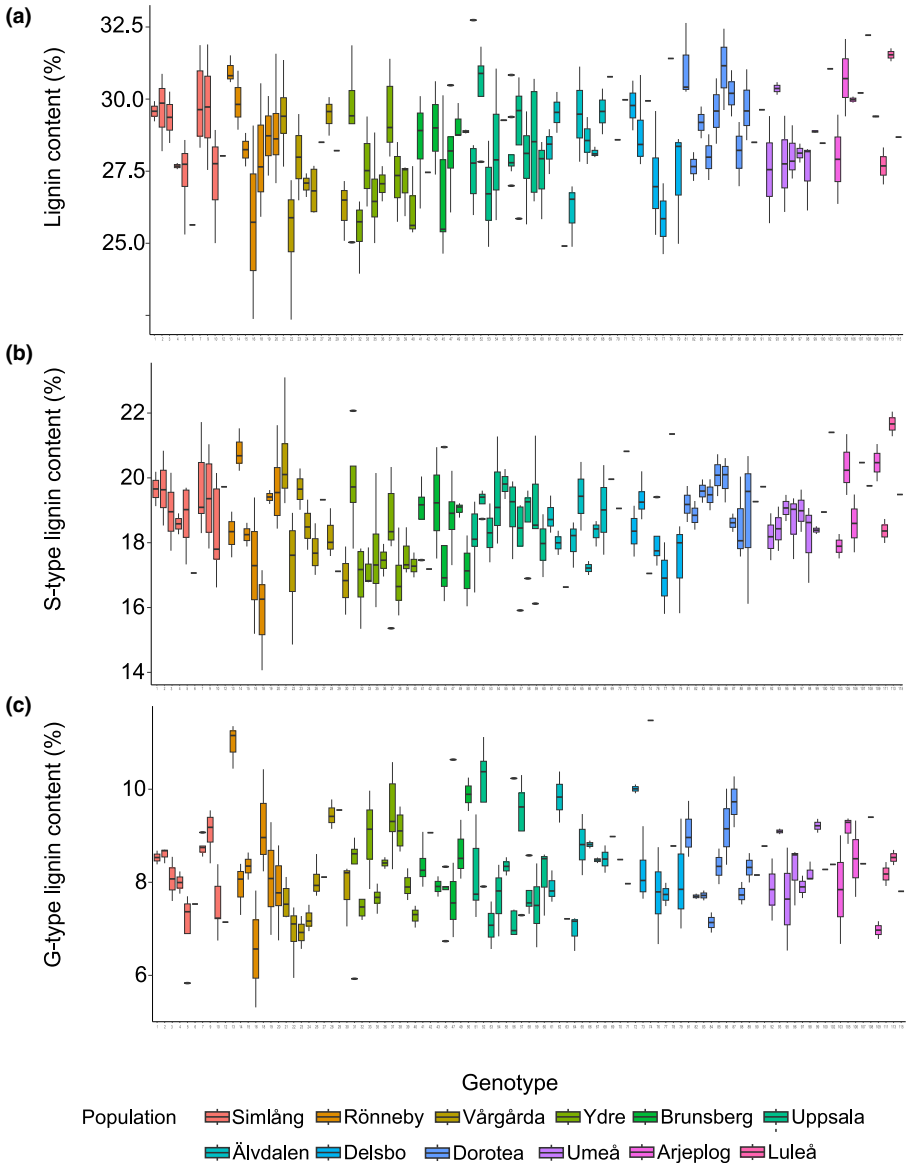


Fig. 1 Wood chemical composition in the Swedish Aspen collection. (a) Relative lignin content. (b) Relative syringyl-type (S) lignin content. (c) Relative guaiacyl-type (G) lignin content. The order of the 99 genotypes follows approximately their latitudinal origin from the south (on the left) to the north (on the right) of Sweden. The colors indicate the geographic origin of the genotypes from 12 different locations in Sweden. Wood chemistry is based on Pyrolysis-GC-MS analysis of mature wood from the base of the stem, whereby the contents are expressed as a percentage of the sum of the area of the peaks corresponding to their respective pyrolytic products relative to the total area of GC peaks. The whiskers extend to 1.5 times the interquartile range. The vertical lines indicate the median for each genotype.

40 cycles, each comprising of denaturation at 95°C for 10s, annealing at 60°C for 10s, and extension at 72°C for 30s and one melting curve ranging from 65°C to 95°C with a step of 0.5°C. All reactions were performed in triplicates in 96-well plates on a C1000Touch™ Thermal Cycler (Bio-Rad) with C_q values acquired with the CFX96™ Maestro software (Bio-Rad). Relative expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method (Livak & Schmittgen, 2001), normalized for each line relative to the average of the WT.

Results

Natural variation in wood chemistry of the Swedish Aspen collection

Wood chemistry was analyzed in 268 SwAsp trees, representing 99 genotypes, by Pyrolysis-GC-MS (Py-GC-MS), which identified the relative contents of the secondary cell wall carbohydrates and the different types of lignin (Fig. 1a–c; Tables 1, S1). The total lignin content varied from 22% to 33% (Fig. 1a). The most abundant lignin subunit was the syringyl-type (S) lignin, which varied from 14% to 23%, followed by guaiacyl-type (G) lignin, which varied from 5.3% to 11% (Fig. 1b,c; Table 1).

Population-wide gene expression in the Swedish Aspen collection

RNA sequencing was performed in differentiating xylem tissues of all SwAsp trees. First, an analysis of the population structure was performed based on the expression of 500 genes with the most variable expression among genotypes. No clear population structure was observed (Fig. S1), which is in agreement with a previous analysis of the population structure in the SwAsp collection (Wang *et al.*, 2018). Broad-sense heritability (H^2) of gene

Table 1 Wood chemical composition in the Swedish Aspen (SwAsp) trees.

Trait	Mean	SD	COV (%)	H^2
Carbohydrates	68.3	1.8	2.7%	0.33
G-type lignin	8.3	1.0	12.5%	0.47
S-type lignin	18.7	1.4	7.6%	0.28
S : G ratio	2.3	0.31	13.5%	0.45
Total lignin	28.4	1.87	6.6%	0.33

Wood chemistry was analyzed in basal stem samples using Pyrolysis-GC-MS. The values for carbohydrates (C), guaiacyl-type (G), syringyl-type (S), and total lignin are relative (%) to the total peak area from the GC-MS. COV, coefficient of variation; H^2, broad-sense heritability.

expression varied between 0 and 0.99 (Fig. 2a) with a mean H^2 of 0.23. Q_{ST}, describing the genetic differentiation among the populations, was low, with a mean value of 0.16 (Fig. 2b). There was a slight positive correlation between median gene expression and broad-sense heritability of gene expression (Pearson $r = 0.13$, df = 34 179, P-value < 2.2e-16), indicating that highly expressed genes are under tighter genetic control than genes with low expression. The correlation between population differentiation (Q_{ST}) and median gene expression was negative (Pearson $r = -0.16$, P-value < 2.2e-16).

An eQTL analysis was performed to detect associations between SNPs and gene expression. Similar to an earlier study (Mähler *et al.*, 2017), hidden confounders, caused by unknown independent variables, were removed from the RNA-seq data. Altogether six confounders were removed without significantly reducing the number of eQTLs (Fig. S2). Using 1 Mbp as a threshold to distinguish between distant and local eQTLs, 292 003 distant (Fig. 2d) and 312 705 local eQTLs (Fig. 2e; Table S2) were found statistically significant (FDR < 0.05). In contrast to local eQTLs, which presumably act through polymorphisms in or around the associated gene, distant eQTLs are likely to influence the expression of the associated gene either by causing a change to the protein-coding sequence of a transcription factor, by affecting expression of a transcription factor or through other mechanisms, such as being within a distantly located enhancer element.

The greatest number of significantly associated SNPs was located either upstream or downstream of coding regions, and less frequently in the intergenic region (Fig. 2c). Earlier eQTL studies have identified eQTL hotspots, that is genomic regions with a higher than expected number of distant acting SNPs (Balmant *et al.*, 2020; Yao *et al.*, 2023; Zhang *et al.*, 2023), although such hotspots are not uniformly reported. Examination of the distribution of distant eQTLs and their target genes across all chromosomes did not reveal any clear hotspots in the current data (Fig. 2f). However, a few genetic loci with a high abundance of *cis* and trans eQTLs were found when considering 1000-bp windows (Fig. 2g; Table S2). Most of these loci were intergenic, requiring further studies to understand their significance (Table S2). An intronic *trans*-eQTL was discovered in FLOWERING LOCUS T 2b (Potra2n10c20839), which together with its associated gene expression provides an interesting source of candidate genes for further studies of wood formation (Table S2).

Whole-transcriptome co-expression network and the lignin subnetwork

The RNA-seq data were used to create a whole-transcriptome co-expression network that consisted of 548 clusters, 1048 575 edges (gene–gene interactions), and 34 623 nodes (genes). The expression data, cluster analysis, co-expression network centrality values, and correlation with lignin traits are presented in Dataset S1.

A lignin subnetwork was constructed based on the expression of *P. tremula* homologs for the lignin-biosynthetic genes from Shi *et al.* (2010) and their first-degree neighbors (Table S3) as well as the correlation with the cell wall chemical components from the Py-GC-MS analysis. Most of the lignin-biosynthetic genes were connected in the lignin subnetwork (Fig. 3a) and showed strong Pearson correlations (Fig. 3b). The lignin traits (total lignin, S-type, and G-type lignin) were also connected in the network (Fig. 3a).

The lignin subnetwork consisted of 14 clusters (Table S3). The lignin-biosynthetic genes were represented in seven different clusters, which might reflect the modularization of lignin biosynthesis. Cluster 17 contained the highest number of lignin-biosynthetic genes and enrichment of GO terms such as phenylpropanoid metabolism and vesicle trafficking (Fig. S3). The clustering pattern suggested connection of *HCT1* (in cluster 81) to metabolism related to pathogen responses, *CAFFEOYL-CoA O-METHYLTRANSFERASE 3* (*CCaOMT3*, in cluster 86) to processes involving reactive oxygen species and abiotic stress, and *4-COUMARATE:CoA LIGASE 3* (*4CL3*, in cluster 87) to circadian rhythms and polysaccharide biosynthesis and metabolism. Somewhat surprisingly, all three *PAL* genes (that are presumably involved in the biosynthesis of several different types of secondary metabolites) were in cluster 202 that was almost exclusively associated with the GO term 'lignin-biosynthetic processes'.

Significant eQTLs were found for nine of the 22 lignin-biosynthetic genes (Fig. S4). *CINNAMATE-4-HYDROXYLASE 2* (*C4H2*; Potra2n19c33285) had both local and distant eQTLs (Fig. S4v), while eQTLs for the other lignin-biosynthetic genes were either local or distant. For instance, a large number of local eQTLs were found for *CAFFEOYL SHIKIMATE ESTERASE 1* (*CSE1*) (Potra2n3c7945; Fig. S4h,w).

The distinct expression patterns of *CAD2* and *CCoAOMT3*

CINNAMYL ALCOHOL DEHYDROGENASE 2 (*CAD2*, Potra2n16c29966) was not part of the lignin subnetwork and did not correlate in expression with the other lignin-biosynthetic genes (Fig. 3a,b). Also, the expression of *CAFFEOYL-COA O-METHYLTRANSFERASE 3* (*CCoAOMT3*, Potra2n8c17885) showed low correlation with the other lignin-biosynthetic genes (Fig. 3b). The expression pattern of these two genes also differed from those of other lignin-biosynthetic genes in the AspWood dataset, which consists of high spatial resolution RNA-seq data from woody tissues of *P. tremula* (Sundell *et al.*, 2017) at PlantGenIE.org (Sundell *et al.*, 2015) (Fig. 3c). Despite the distinct co-expression pattern of these genes, *CAD2* correlated

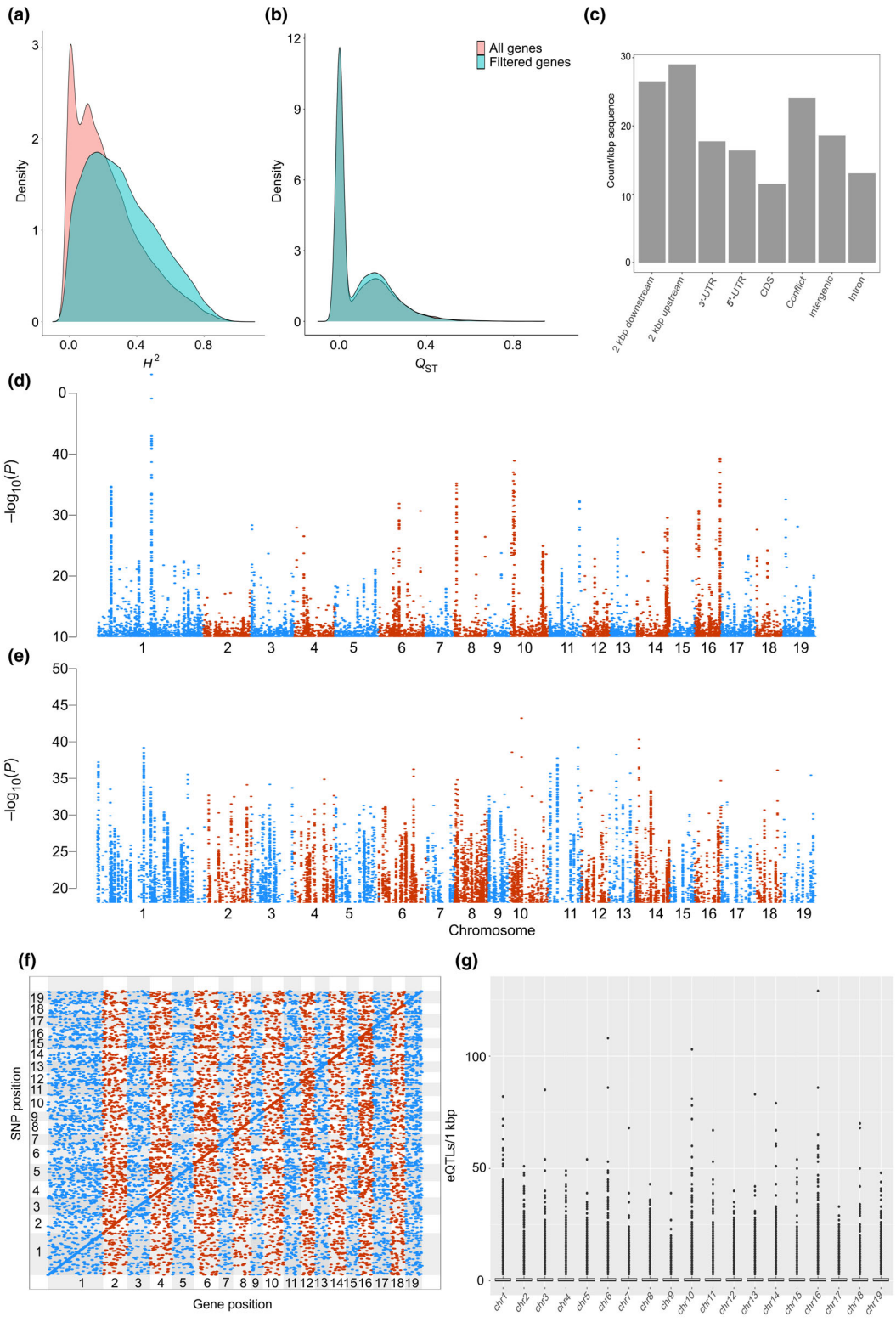
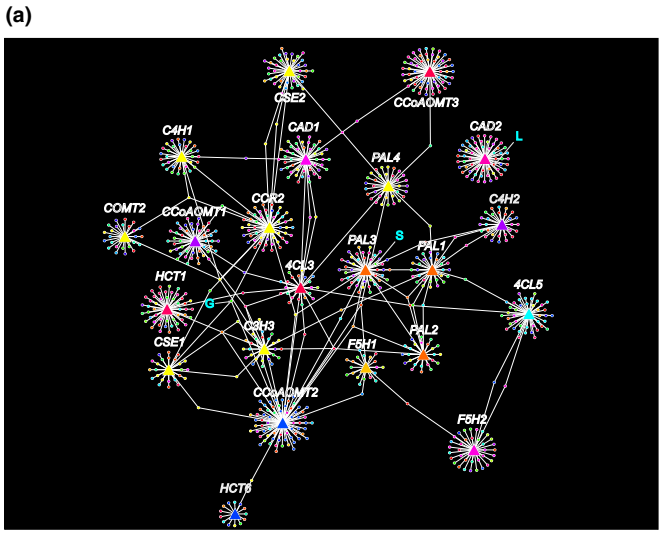


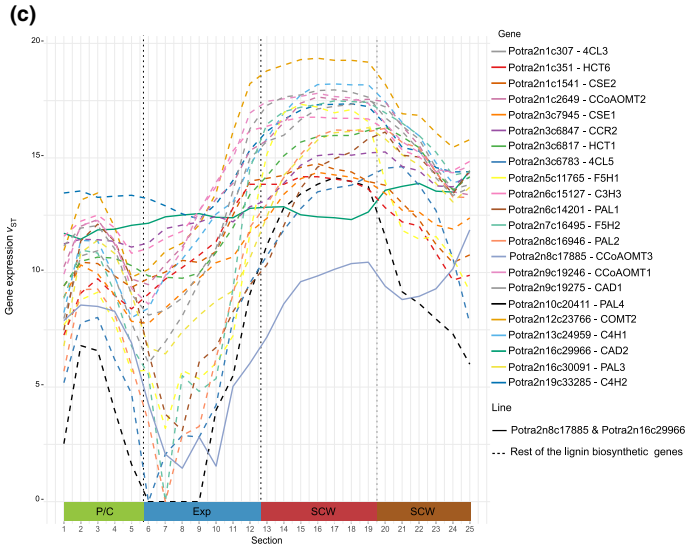
Fig. 2 Overview of the expression quantitative trait loci (eQTL) analysis in the Swedish Aspen collection. (a, b) Distribution of gene expression heritability (a) and quantitative genetic differentiation Q_{ST} (b). The red color represents all expressed genes, and the blue color genes were filtered out due to low variation or low expression. (c) Genomic context for the significant (false discovery rate (FDR) < 0.05) eQTLs. The number of eQTLs per context category was normalized for the feature length. The conflict group consists of eQTLs assigned into more than one category. (d, e) Manhattan plot for distant eQTLs (d) and local eQTLs (e). The statistical significance ($-\log_{10}(P)$-value) is shown for each eQTL according to its genomic position. (f) The genomic positions of the significant (FDR < 0.05) distant eQTLs. Numbers indicate the number of the chromosome. (g) The number of significant (FDR < 0.05) distant eQTLs within genomic blocks of 1000 bp along the different chromosomes.

positively with the S-type and total lignin content of the wood while *CCoAOMT3* correlated negatively with total lignin content (Fig. 3b), supporting their regulatory roles in lignin biosynthesis.

CAD2 and *CCoAOMT3* had also distinct co-expression patterns with the transcription factors in the lignin subnetwork (Table S4). *CAD2* was co-expressed with four transcription factors, out of which *RGA-like 2* (Potra2n17c30832) was the only one with high expression in the AspWood dataset. *CCoAOMT3* was co-expressed only with *NIN-like* transcription factor (Potra2n4c10214) which is expressed in the phloem and during late xylem maturation in the AspWood dataset.

Based on the analysis of mapped transcription factor binding motifs obtained from the JASPAR CORE 2020 plant database





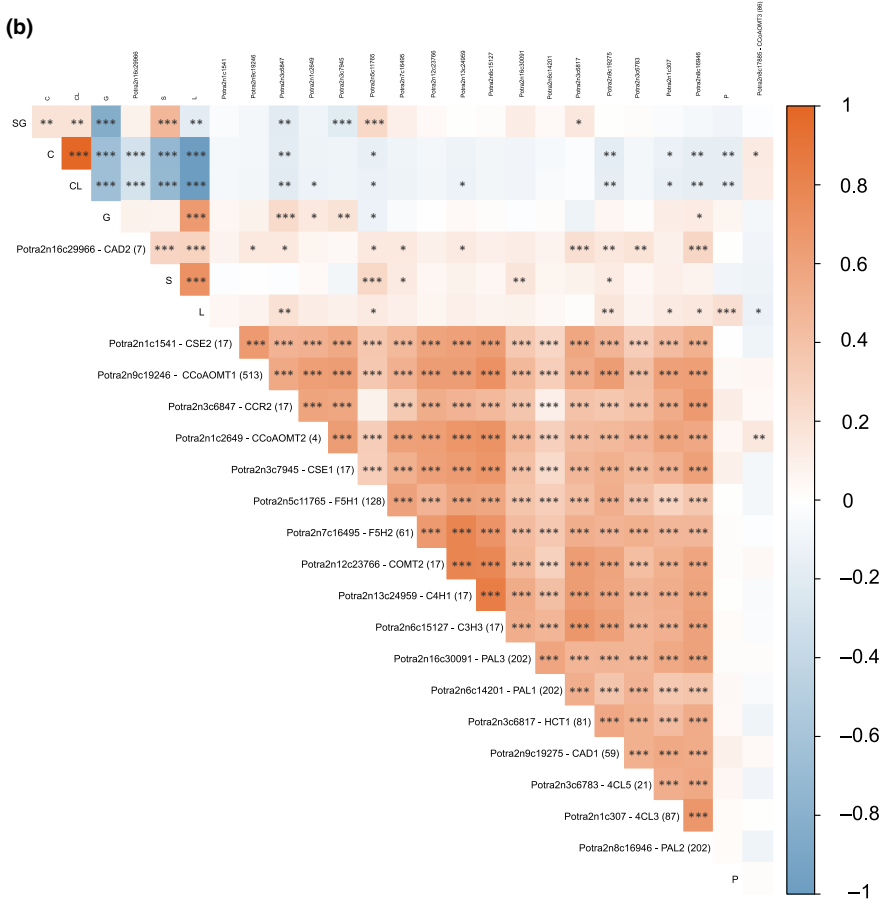


Fig. 3 Expression of lignin-biosynthetic genes in the Swedish Aspen collection. (a) The lignin subnetwork connecting lignin traits and the expression of lignin-biosynthetic genes and their first-degree neighbors. Each hub represents a lignin-biosynthetic gene (represented as a triangle) and its first-degree neighbor (represented as circles) based on the whole-transcriptome co-expression network. Node color is assigned based on the infomap clustering (Supporting Information Dataset S1). The lignin traits were derived from the Pyrolysis-GC-MS analysis. L, relative total lignin content; S, relative syringyl lignin content; G, relative guaiacyl lignin content. The *Populus tremula* lignin-biosynthetic genes were identified on the basis of the highest BLASTN *e*-value to the *Populus trichocarpa* lignin-biosynthetic genes listed in Shi *et al*. (2010). (b) Heatmap representation of the correlation between the Pyrolysis-GC-MS traits and the expression of the lignin-biosynthetic genes. The colors correspond to Pearson correlation values. Asterisks indicate the statistical significance of the Pearson correlation coefficients using *t*-distribution; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. C, relative total carbohydrate content; G, relative guaiacyl lignin content; S, relative syringyl lignin content; P, relative content of generic benzene derivatives, most probably originated from lignin in plants; L, relative total lignin content; S : G, syringyl and guaiacyl lignin ratio; CL, carbohydrate and lignin ratio. (c) The expression of the lignin-biosynthetic genes in aspen stem tissues. The data on tangential cryosections (1–25) across the cambial tissues of aspen (Tree 1) are obtained from the AspWood dataset (Sundell *et al*., 2017) in PlantGenIE.org (Sundell *et al*., 2015). Gene expression values are VST normalized. CD, cell death; Exp, expanding xylem; P/C, phloem; SCW, secondary cell wall formation.

(Fornes *et al*., 2020), there were no clear differences in the presence or absence of motifs in promoter regions of *CAD2* compared with the promoters of the other lignin-biosynthetic genes. However, *CCoAOMT3* had a profile that differed from the other lignin-biosynthetic genes (Fig. S5).

Transcriptional regulation of the lignin-biosynthetic pathway

The lignin subnetwork allowed elucidation of the transcriptional regulation of the lignin-biosynthetic genes. Eighty-nine transcription factors, enriched with WD and MYB-like domain-containing factors, were identified among the 850 genes of the lignin subnetwork (Table S4). These included transcription factors, such as the homologs for the Arabidopsis *MYB42* and *MYB85*, known to control the expression of the lignin-biosynthetic genes (Zhong *et al*., 2008; Geng *et al*., 2020), but mostly transcription factors not previously connected to transcriptional regulation of lignin biosynthesis. One of the transcription factors in the lignin subnetwork was Potra2n1c1657, annotated as *HB5* (*HB5 aka POPCORONA* in *P. trichocarpa*; Du *et al*., 2011), which belongs to the Class III homeodomain-leucine zipper family (Fig. 4a,b). *HB5* was co-expressed with the lignin-biosynthetic *PAL3* (Potra2n16c30091), which in turn co-varied with the S-type lignin content (Fig. 3a).

HB5 is expressed in the cambial region and differentiating xylem elements according to the AspWood dataset (Fig. 4c). The impact of HB5 on lignin content was confirmed in two *Populus tremula* × *P. tremuloides* 'T89' RNAi lines 5AA and 5AB with decreased expression of both *HB5* and its paralog, *HB6* (Potra2n3c8027) (Fig. 4d). Both G- and S-type lignin contents were significantly reduced in the RNAi lines compared with the T89 WT control (Fig. 4e). The plant height and stem diameter of the RNAi lines were increased (Figs 4f,g, S6a) while wood density (Fig. 4h) and anatomy (Fig. S6b,c) were similar to that of the control plants. As expected on the basis of co-expression of *PAL3* with *HB5*, the expression of *PAL3* was lower in the transgenic *HB5* RNAi lines when compared to control even though statistically significant in only one of the lines (Fig. S7a). Accordingly, a binding motif was present for AtHB-15, the closest Arabidopsis homolog of the *Populus HB5*, in the *PAL3* 2-kb promoter region (Fig. S7b). These results demonstrate that HB5 is involved in regulating lignin content, and hence provide functional validation for the lignin subnetwork.

Genome-wide association analysis of wood chemistry in the Swedish Aspen collection

GWAS was performed for eight wood chemical composition traits (as listed in Table S1) in 99 SwAsp genotypes (Table S5). The strongest associations were obtained for G- and S-type lignin (Table 2).

For S-type lignin, the highest association was observed for a SNP in a UDP-glucosyl transferase family protein Potra2n1c2130 (Tables 2, S5; Fig. 5a) but which was not expressed in the developing xylem according to the AspWood dataset and therefore not analyzed further. The second highest association was for SNPs in a galactosyl transferase family protein Potra2n1c3762, annotated as *GT31_32* in Kumar *et al*. (2019) (Tables 2, S5). *GT31_32* is homolog of the Arabidopsis *GLYCOSYLTRANSFERASE GT31 A* (*GALT31A; AT1G32930*; Fig. 5b) which has been implicated in glycosylation of arabinogalactan proteins (Geshi *et al*., 2013). A role of *GT31_32* in arabinogalactan homeostasis was supported by co-expression with four homologs of the Arabidopsis *FASCICLIN-LIKE ARABINOGALACTAN PROTEIN 12* (*FLA12*) in the whole-transcriptome co-expression network (Dataset S1). In the AspWood dataset, *GT31_32* is highly expressed in the cambial region and expanding xylem as well as during late xylem maturation (Fig. 5c). The SNP chr1_48493303A_G in *GT31_32* resulted in an exonic, synonymous mutation (Table 2). The genotypic variants for the homozygous and heterozygous alleles of chr1_48493303A_G showed statistically significant differences in S-type lignin content (Fig. 5d), but not in the expression of *GT31_32* (Fig. 5e), which suggests that GT31_32 does not influence S-type lignin accumulation through changes in its expression. However, the genomic region of *GT31_32* contained several local eQTLs (Fig. 5f). Furthermore, *GT31_32* was co-expressed with *CCoAOMT2*, and shared 19 co-expressed genes with *CCoAOMT2* (Potra2n1c2649) in the lignin subnetwork (Table S3), which altogether supports the importance of *GT31_32* expression in lignin accumulation. Notably, the expression of *GT31_32* correlated negatively with the expression of *CCoAOMT2* (Dataset S1).

(a)

Gene Name	*P. trichocarpa*	*P. tremula*	*P. tremula* × *P. treumloides* (T89)	*Arabidopsis thaliana*
HB1	Potri.009G014500	Potra2n9c19989	Potrx053501g00010	AT5G60690, REVOLUTA
HB2	Potri.004G211300	Potra2n4c10264	Potrx062371g00010	AT5G60690, REVOLUTA
HB3	Potri.011G098300	Potra2n11c23064	Potrx053214g00010	AT2G34710, PHABULOSA, AtHB-14
HB4	Potri.001G372300	Potra2n1c3263	Potrx065319g00010	AT2G34710, PHABULOSA, AtHB-14
HB5	Potri.001G188800	Potra2n1c1657	Potrx046843g13917	AT1G52150, CORONA, AtHb-15
HB6	Potri.003G050100	Potra2n3c8027	Potrx065083g25217	AT1G52150, CORONA, AtHb-15
HB7	Potri.018G045100	Potra2n18c32866 Potra2n18c32867		AT4G32880, AtHB-8
HB8	Potri.006G237500	Potra2n6c13216	Potrx057596g00010	AT4G32880, AtHB-8

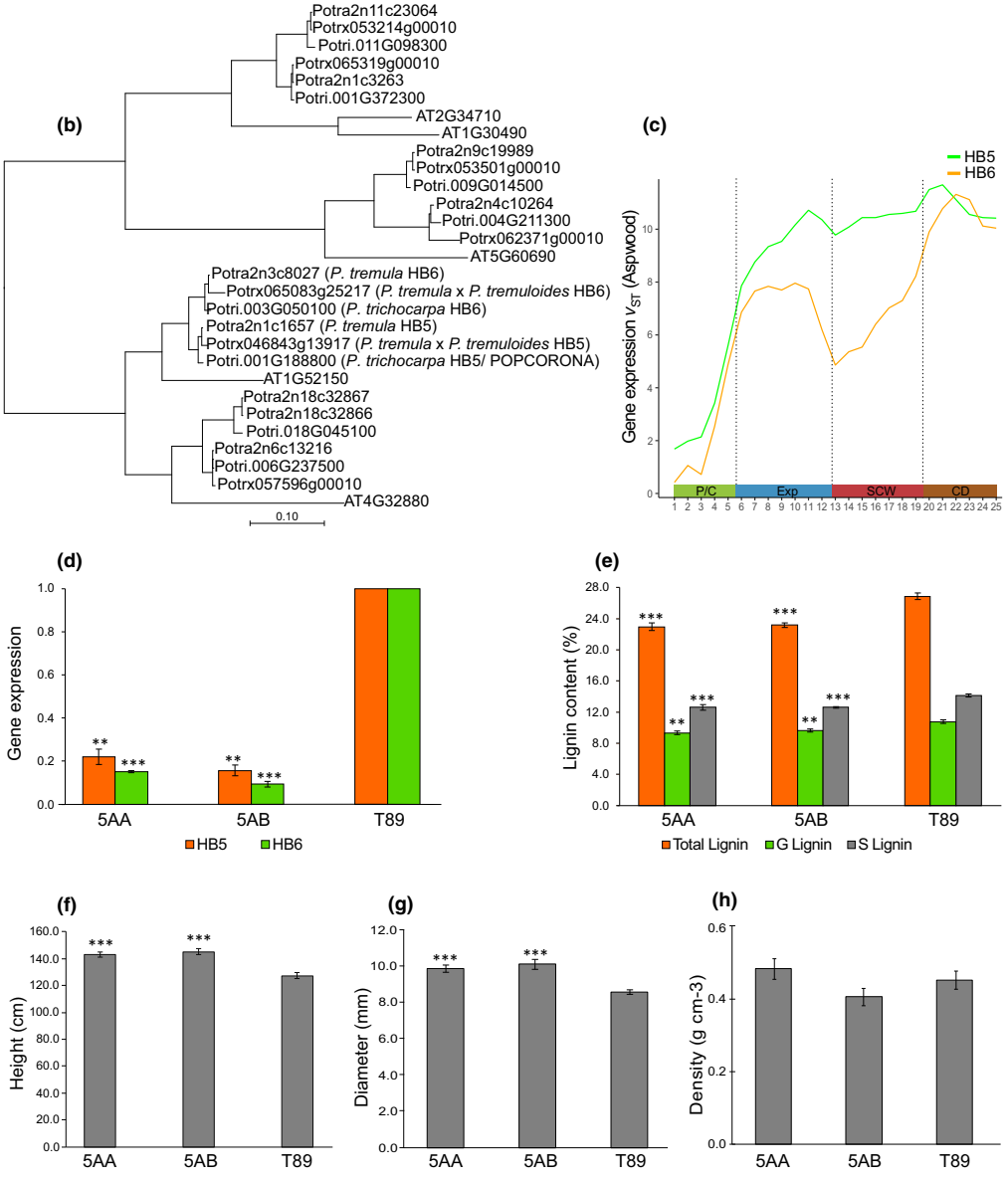


Fig. 4 Function of the HOMEOBOX PROTEIN 5 (HB5) in the regulation of lignin biosynthesis. (a) Class III homeodomain-leucine zipper (HD-Zip III) gene family in *Arabidopsis thaliana*, *Populus tremula*, *Populus trichocarpa*, and *P. tremula* × *P. tremuloides*. The sequences were retrieved from PlantGenIE.org (Sundell *et al*., 2015). (b) Phylogeny of the HD-Zip III gene family reconstructed using iTOL/v6 (https://itol.embl.de/). (c) Expression profiles of *HB5* and *HB6* in aspen stem tissues corresponding to phloem (P/C), expanding xylem (Exp), secondary cell wall (SCW) formation, and cell death (CD). Gene expression values (VST normalized) represent Tree1 in the aspen wood (AspWood) dataset (Sundell *et al*., 2017) from PlantGenIE.org (Sundell *et al*., 2015). (d) Relative expression of *HB5* and *HB6* in T89 wild-type (WT) and *HB5* RNAi lines 5AA and 5AB. The data represent average $2^{-\Delta\Delta Ct}$ values from quantitative reverse transcription polymerase chain reaction analysis. $n = 4$–5. (e) Total lignin, G-type lignin, and S-type lignin content in T89 WT and *HB5* RNAi lines 5AA and 5AB. The data are obtained with Pyrolysis-GC-MS. (f–h) Tree height (f), stem diameter (g), and wood density (h) at the base of the stem in two *HB5* 5AA and 5AB RNAi lines and T89 WT. The experiment with the transgenic lines was repeated twice. Asterisks indicate significant differences from the T89 WT at **, $P < 0.01$; or ***, $P < 0.001$ according to a *t*-test. Error bars indicate SD.

Table 2 Top-ranked genes in genome-wide association study (GWAS) for wood chemical composition in the Swedish Aspen (SwAsp) collection.

Trait	Locus	SNP	No SNPs in locus	*P*-value	*q*-Value	PVE	Feature	Arabidopsis gene	Arabidopsis description
G-type lignin	Potra2n11c22479	chr11_1139060C_G chr11_1139068T_A chr11_1139090G_A	3	9.29E-09	0.031	0.29	Upstream Upstream Upstream	AT5G44520	RPI-like, NagB/RpiA/CoA transferase-like superfamily protein
	Potra2n3c6909 & Potra2n3c6910	chr3_4092633G_C	1	2.26E-08	0.037	0.28	Intergenic		
S-type lignin	Potra2n1c2130	chr1_25487896T_A	1	1.47E-08	0.136	0.29	Exonic, nonsynonomous	AT3G46660	UDP-glucosyl transferase 76E12
	Potra2n1c3762	chr1_48493303A_G chr1_48493246T_G chr1_48493255T_C	3	4.16E-08	0.136	0.28	Exonic, synonomous Exonic, synonomous Exonic, synonomous	AT1G32930	GALT31A, Galactosyltransferase family protein

PVE, phenotypic variation explained; SNP, the genomic location of the single-nucleotide polymorphism.

The three most significant SNPs ($q < 0.05$) for G-type lignin were located within 1500-bp upstream of *RIBOSE-5-PHOSPHATE ISOMERASE-LIKE* (*RPI-like*) Potra2n11c22479 (Fig. 6a; Tables 2, S5), the closest homolog to Arabidopsis RPI4 (At5g44520; Fig. 6b). The aspen *RPI-like* is highly expressed in the phloem, cambium, and expanding xylem according to the AspWood dataset (Fig. 6c). The most significant SNP (chr11_1139060C_G) was associated with variation in G-lignin content (Fig. 6d), but not *RPI-like* expression (Fig. 6e) even though multiple local SNPs were found upstream of the *RPI-like* in the eQTL analysis (Fig. 6f). The association between SNP chr11_1139060C_G and G-lignin content was verified also in an independent aspen population (UmAsp; Fig. S8).

Discussion

Natural variation in lignin chemistry in Swedish aspen

We observed a range of 22 to 33% variation for the relative lignin content of woody tissues in a geographically representative collection of aspen trees from Sweden (Fig. 1a; Table S1). This variation is similar to earlier observations in other *Populus* species (Studer *et al.*, 2011; Guerra *et al.*, 2013; Porth *et al.*, 2013; Muchero *et al.*, 2015; Fahrenkrog *et al.*, 2017). The S : G ratio varied between 1.4 and 3.2 (Table S1), which exceeds the magnitude of variation reported previously for *P. nigra* and *P. deltoides* (Guerra *et al.*, 2013; Fahrenkrog *et al.*, 2017). In view of these results, and given the importance of S : G-lignin ratio on the recalcitrance of woody biomass (Studer *et al.*, 2011; Yoo *et al.*, 2017; Anderson *et al.*, 2019), it is reasonable to consider the lignin subunit composition as a promising trait for any *Populus* tree improvement initiative.

Variation in a quantitative trait such as lignin content is crucial for adaptation to the surrounding environment and ultimately the survival of the species (Crivellaro *et al.*, 2022). One such environmental factor is the day length, which causes strong latitudinal clines on many traits in Sweden (Wang *et al.*, 2018; Lihavainen *et al.*, 2023). However, there were no latitudinal clines in lignin content or composition within the SwAsp population (Fig. 1). Variation in lignin content of SwAsp trees should therefore serve other purposes, such as responses to light, circadian clock, defense against pathogens, and abiotic as well as mechanical stresses that are all known to influence lignin biosynthesis (Chantreau & Tuominen, 2022).

Unique aspects of the transcriptional regulation of lignin biosynthesis

The extent of natural variation in the expression of the lignin-biosynthetic genes remains poorly understood. The eQTL analysis revealed significant variation ($P < 5 \times 10^{-8}$) in the expression of almost half of the lignin-biosynthetic genes (Fig. S4). However, the variation did not seem to influence lignin content since the genes were either poorly or not at all associated with lignin traits in the GWAS (Table S3).

Most of the lignin-biosynthetic genes, except for *CAD2* and *CCoAOMT3*, correlated in expression within the SwAsp collection and had similar spatial expression patterns (Fig. 3). Interestingly, *CAD2* and *CCoAOMT3* had the most distinct increase in expression in the AspWood dataset in the last sample of the series (Fig. 3c), which represents gene expression in rays since the fibers are already dead at this point (Sundell *et al.*, 2017). Therefore, these genes might be regulated by factors derived from the sap. Ray expression has been shown for members of both *CAD* and *CCoAOMT* families (Feuillet *et al.*, 1995; Chen *et al.*, 2000). However, the importance of CAD2 and CCoAOMT3 in lignification has remained somewhat unclear especially since Crispr Cas9 knockout in either of them does not influence lignin content or composition (Sulis *et al.*, 2023).

CAD2 was earlier defined as SINAPYL ALCOHOL DEHYDROGENASE (SAD) having a sinapyl aldehyde specific enzymatic activity (Li *et al.*, 2001). More recently, CAD2 was shown to accept both coniferyl aldehyde and sinapyl aldehyde as the substrate even though sinapyl aldehyde was the preferred substrate on the basis of lower Michaelis constant (Km) values (Chao *et al.*, 2014; Wang *et al.*, 2014). Our analysis identified positive correlation between the expression of *CAD2* and the S-type lignin content (Fig. 3b), which supports the preferential function of CAD2 toward S-type lignin biosynthesis. However, functional studies in tobacco (Barakate *et al.*, 2011) or poplar (Sulis *et al.*, 2023) have not supported the function of CAD2 on S-type lignin biosynthesis.

New Phytologist

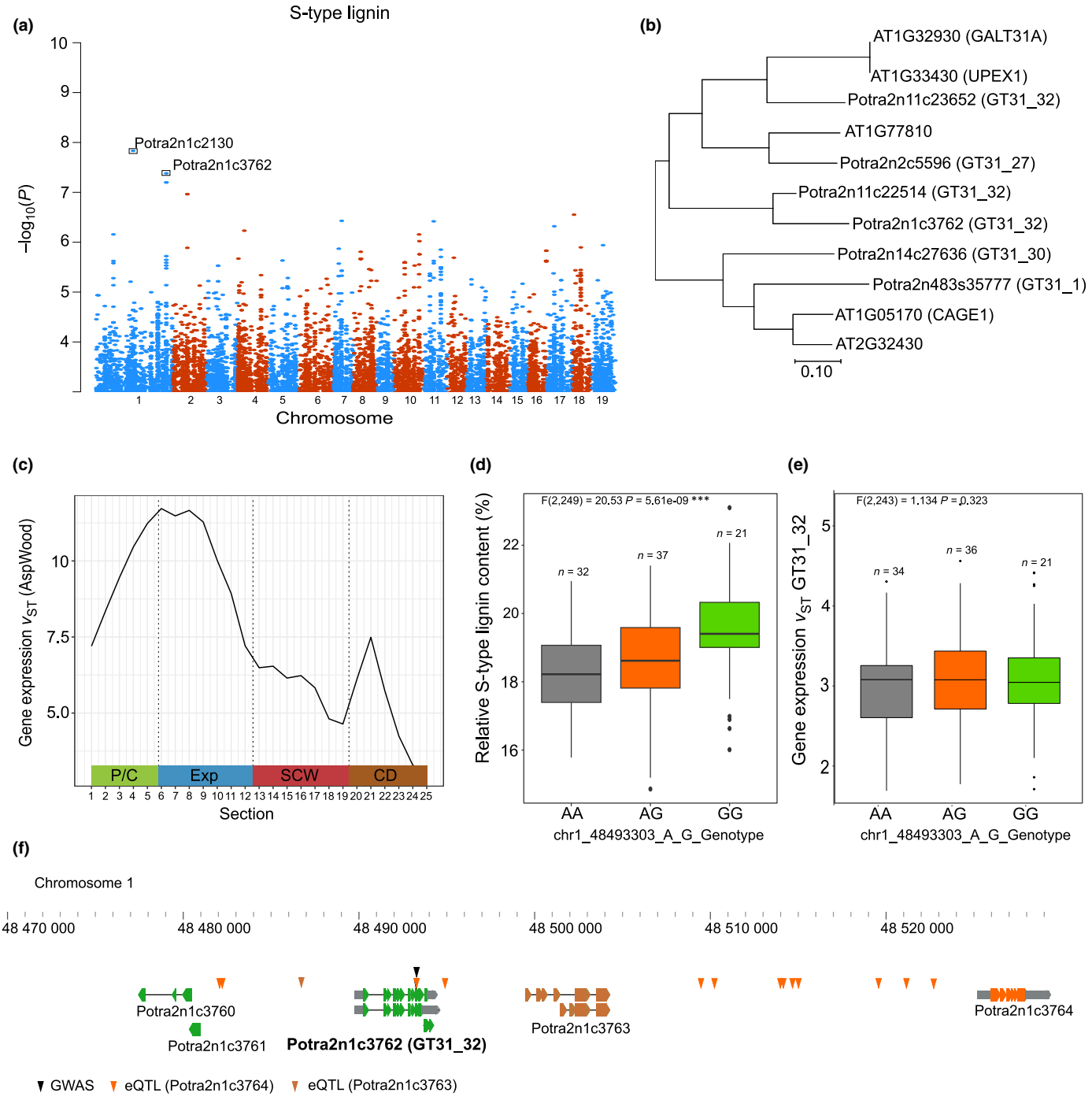


Fig. 5 Overview of the genome-wide association study results for the S-type lignin in the Swedish Aspen (SwAsp) collection. (a) Manhattan plot displaying genome-wide associations for S-type lignin in the SwAsp collection. The statistical significance of the association ($-\log_{10}(P)$-value) is shown for each single-nucleotide polymorphism (SNP) according to its genomic position. The top-ranked SNPs are enclosed by a square and indicated with the corresponding gene models. (b) Phylogeny of galactosyltransferase GT31_32 in *Populus tremula* and *Arabidopsis thaliana*. (c) Expression profile of the galactosyltransferase *GT31_32* (Potra2n1c3762) in sections across the cambial region in the aspen stem. Gene expression values are from Tree 1 of the aspen wood (AspWood) dataset (Sundell *et al*., 2017) from PlantGenIE.org (Sundell *et al*., 2015). P/C, phloem; SCW, secondary cell wall; CD, cell death zone. (d, e) Relative S-type lignin content (d) and expression of *GT31_32* (e) in the three genotypic groups of the SNP chr1_48493303A_G in *GT31_32* in SwAsp collection. The relative S-type lignin content was estimated by Pyrolysis-GC-MS. Gene expression values are VST normalized. Analysis of variance *F*-ratios and *P*-values are shown. ***, $P < 0.0001$. (f) Representation of the genomic region containing SNP chr1_48493303A_G_for *GT31_32* in chromosome 1. The black downward-pointing arrowhead indicates the location of the SNP chr1_48493303A_G from GWAS. The brown arrowhead indicates a local expression quantitative trait locus (eQTL) for *GT31_32* (Potra2n1c3762). The orange arrowheads indicate local eQTLs for Potra2n1c3764.

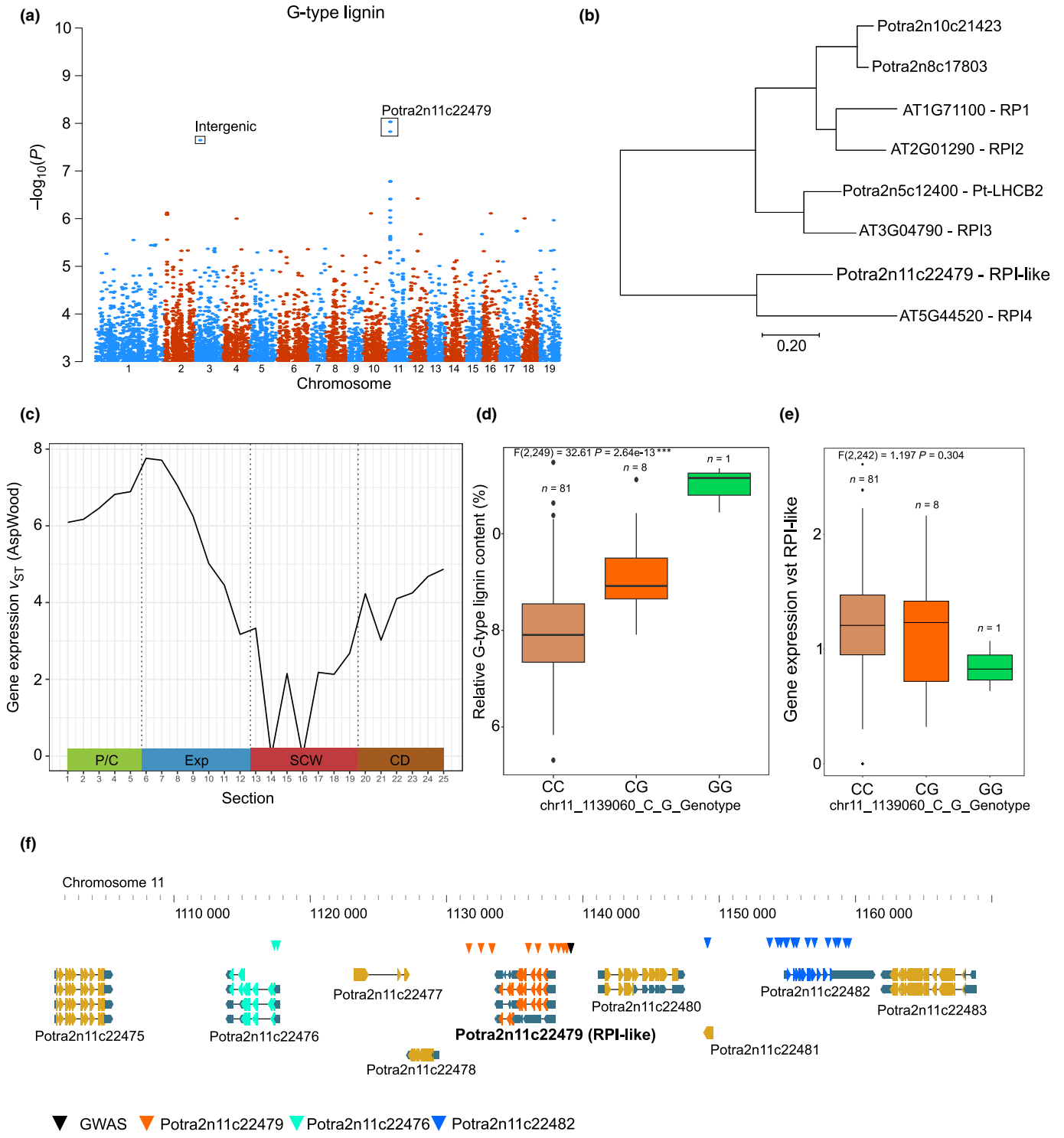


Fig. 6 Overview of the genome-wide association study results for the G-type lignin in the Swedish Aspen (SwAsp) collection. (a) Manhattan plot displaying associations for G-type lignin in the SwAsp collection. The statistical significance ($-\log_{10}(P)$-value) is shown for each individual variant according to its genomic position. The top-ranked single-nucleotide polymorphisms (SNPs) are enclosed by a square and indicated with the corresponding gene models. (b) Phylogeny of the RIBOSE-5-PHOSPHATE ISOMERASE (RPI) gene family in *Populus tremula* and *Arabidopsis thaliana*. (c) The expression profile of *P. tremula RPI-like* (Potra2n11c22479) in sections across the cambial region in the aspen stem. Gene expression values are from Tree 1 of the aspen wood (AspWood) dataset (Sundell *et al*., 2017) from PlantGenIE.org (Sundell *et al*., 2015). CD, cell death zone; P/C, phloem; SCW, secondary cell wall. (d, e) The relative G-type lignin content (d) and the expression of the *RPI-like* (e) in the three genotypic groups of the SNP chr11_1139060C_G in *RPI-like* in SwAsp collection. The relative G-type lignin content was estimated by Pyrolysis-GC-MS. Gene expression values are VST normalized. Analysis of variance *F*-ratios and *P*-values are shown. ***, $P < 0.0001$. (f) Representation of the genomic region containing the SNP chr11_1139060C_G in the *RPI-like* in chromosome 11. Arrowheads depict local eQTLs, and the color indicates the gene with which the SNP associates.

CCoAOMT3 had a distinct profile for transcription factor binding motifs among the lignin-biosynthetic genes (Fig. S5). A NIN-like transcription factor, *NLP1a*, was co-expressed exclusively with *CCoAOMT3* in the lignin network (Table S4). The *Populus* NLP1a is the closest homolog of Arabidopsis NLP2, which regulates nitrogen assimilation and metabolism (Liu *et al.*, 2022). A link between lignin and nitrogen availability is well known (Pitre *et al.*, 2007), and the expression of *CCoAOMT3* has also been linked with nitrogen availability (Cooke *et al.*, 2003; Zhao *et al.*, 2022). It is therefore possible that the variation in lignin content in response to nitrogen availability is mediated through changes in *CCoAOMT3* expression. Interestingly, the expression of *CCoAOMT3* correlated negatively with lignin content (Fig. 3b), suggesting that CCoAOMT3 might have an unknown function in suppressing the accumulation of lignin.

The function of *Populus* AtHB-15 homologs in lignification

Members of the Arabidopsis HD-ZIP III family members, such as PHAVOLUTA, PHABULOSA, REVOLUTA, AtHB-15/CORONA, and AtHB-8, are linked to the regulation of meristem activities and organ polarity. In the current study, evidence was obtained on the function of HB5, the *Populus* homolog of AtHB-15, in lignification. *HB5* was co-expressed with *PAL3* in the SwAsp population, and suppression of *HB5* together with its paralog *HB6 by RNAi* resulted in lower lignin content of transgenic trees, supporting their role in stimulating lignin biosynthesis through *PAL3* (Figs 4, S7). Even though evidence was obtained on a direct function of HB5 on *PAL3* expression (Fig. S7b), it is possible that the effect of HB5 is indirect through the negative effect it has on the radial growth of the stem. The negative effect of HB5 on radial growth of the stem is consistent with earlier reports on AtHB-15 in Arabidopsis (Kim *et al.*, 2005; Wei *et al.*, 2023) and on the AtHB-15 homologs POPCORONA (HB5) and Pt.AtHB.11 (HB6) in *Populus* (Du *et al.*, 2011). The role of the AtHB-15 homologs and the other members of the HD-ZIP III family members has however largely been overlooked in xylem differentiation.

Arabinogalactans proteins and S-type lignin accumulation

One of the top-ranked genes in the GWAS analysis of S-type lignin was *GT31_32* which encodes a protein that is likely involved in the glycosylation of arabinogalactan proteins (Geshi *et al.*, 2013) (Fig. 5). *GT31_32* was co-expressed with both the lignin-biosynthetic *CCoAOMT2* and several *FASCICLIN-LIKE ARABINOGALACTAN* (*FLA*) genes (Dataset S1), supporting interaction between arabinogalactan proteins and lignin in the secondary cell wall context. *Populus* species have 35 *FLA* genes (Zang *et al.*, 2015), which has hampered their functional characterization. However, it is well known that both the expression of *FLA* genes and the accumulation of arabinogalactan proteins are stimulated during tension wood formation, coinciding with suppressed accumulation of lignin (Pilate *et al.*, 2004; Lin *et al.*, 2022). The negative correlation between the expression of *FLA* genes and *CCoAOMT2* in our data (Dataset S1) supports the mutually exclusive accumulation of the arabinogalactan proteins and lignin. On the basis of these results, we hypothesize that the function of GT31_32 is related to S-type lignin accumulation indirectly through its regulatory function on the arabinogalactan proteins in situations when lignin biosynthesis is suppressed, such as during tension wood formation, in favor of the biosynthesis of arabinogalactan proteins and cellulose. This would also mean that there is variation among the trees in their response to mechanical stimulus resulting in tension wood formation.

Pentose phosphate pathway and G-type lignin biosynthesis

Variation in the *RPI-like* (Potra2n11c22479) sequence was significantly associated with the content of G-lignin in two independent aspen populations (Figs 6a, S8). Arabidopsis has four ribose-5-phosphate isomerases; *RPI1* (At1g71100), *RPI2* (At2g01290), *RPI3* (At3g04790), and *RPI4* (At5g44520), which belong to three different clades (Xiong *et al.*, 2009). RPIs are involved in the pentose phosphate pathway that feeds into the biosynthesis pathways of both cellulose and lignin (Stincone *et al.*, 2015). Accordingly, a mutation in *RPI1* reduced the level of cellulosic glucose in Arabidopsis seedlings (Howles *et al.*, 2006). It is therefore possible that the observed association between *RPI-like* and G-lignin content is indirect and that the primary cause for the observed variation is through the impact of the *RPI-like* variant on cellulose biosynthesis. However, a mutation in another pentose phosphate pathway gene, *TRANSALDOLASE 2* (*TRA2*), resulted in a 15% reduction in lignin content and an increased S : G ratio in Arabidopsis (Vanholme *et al.*, 2012). Possible homologs of the Arabidopsis *TRA2* were also found associated with H-lignin (Bryant *et al.*, 2023). Altogether, these results support the effect of pentose phosphate pathway and hence also RPI-like primarily on the lignin pathway and in particular the lignin composition.

In conclusion, a systems genetic study was performed on lignin biosynthesis and its regulation in the SwAsp collection. The analysis revealed associations for complex traits such as G- and S-type lignin, including an association between a *RPI-like* gene and G-type lignin content that was also validated in the independent Umeå aspen (UmAsp) collection. New insights were gained into the expression and regulation of lignin-biosynthetic genes, including distinct features of *CAD2*, *CCoAOMT2*, and *CCoAOMT3* expression. While *CAD2* contributes preferentially to S-type lignification, *CCoAOMT2* and *CCoAOMT3* seem to control lignification in response to specific external conditions, such as tension wood formation and nitrogen status, respectively. Taken together, this work demonstrates the value of systems genetics in studying the genetic architecture of complex traits such as lignin content and composition in woody tissues.

Acknowledgements

We thank Sara Ölmelid for help in collecting the SwAsp material and preparing the wood samples for the analyses, Sara Westman for help with the GWAS, Hui Li for help with the eQTL analysis,

SweTree Technologies for generating the transgenic *HB5* RNAi lines, the UPSC bioinformatics facility for help with bioinformatics, the facilities and technical assistance of the Umeå Plant Science Centre (UPSC) Microscopy Facility, UPSC Biopolymer Analytical Platform (supported by Bio4Energy and T4F) and its manager, Junko Takahashi-Schmidt, for the Pyrolysis-GC-MS measurement, and Skogforsk at Ekebo and Sävar for hosting the SwAsp and UmAsp common gardens. The computations were enabled by resources in the project (UPPMAX SNIC 2019/8-324) provided by Uppsala University at UPPMAX. This work was supported by grants from Formas (2018-01611 and 2018-01381), the Knut and Alice Wallenberg Foundation (2016.0341 and 2016.0352), the Swedish Governmental Agency for Innovation Systems VINNOVA (2016-00504), the strategic research initiatives Bio4Energy (www.bio4energy.se) and Trees for the Future (T4F), and Fundação para a Ciência e Tecnologia, Portugal through CEEC/IND/00175/2017, GREEN-IT (UIDB/04551/2020 and UIDP/04551/2020) and LS4FUTURE (LA/P/0087/2020).

Competing interests

ON, NRS, and HT are shareholders of Woodheads AB.

Author contributions

CG and HT planned and designed the research. CG coordinated the collection of the material, RNA extractions, and RNA-seq. ON contributed to the RNA-seq efforts. ML analyzed the RNA-seq data and performed the GWAS, eQTL, and network analyses. KMR assisted in the GWAS analysis. TAK performed the JASPAR motif analysis. AM and SC analyzed the *HB5* RNAi lines. NRS supervised the bioinformatic analyses. ML and HT wrote the manuscript with the help of all co-authors.

ORCID

Carolin Grones https://orcid.org/0000-0002-8962-3778
Ana Milhinhos https://orcid.org/0000-0002-5699-0010
Ove Nilsson https://orcid.org/0000-0002-1033-1909
Kathryn M. Robinson https://orcid.org/0000-0002-5249-604X
Nathaniel R. Street https://orcid.org/0000-0001-6031-005X
Hannele Tuominen https://orcid.org/0000-0002-4949-3702

Data availability

The RNA-sequencing data of the SwAsp clones used in the analysis are available at NCBI SRA resource as BioProject PRJNA297202 (SRA: SRP065057).

References

Akyol T. 2019. *RVENN: set operations for many sets.* R package v.1.1.0.

Anderson EM, Stone ML, Katahira R, Reed M, Muchero W, Ramirez KJ, Beckham GT, Román-Leshkov Y. 2019. Differences in S/G ratio in natural poplar variants do not predict catalytic depolymerization monomer yields. *Nature Communications* **10**: 2033.

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME suite. *Nucleic Acids Research* **43**: W39–W49.

Balmant KM, Noble JD, Alves FC, Dervinis C, Conde D, Schmidt HW, Vazquez AI, Barbazuk WB, de Los CG, Resende MF *et al.* 2020. Xylem systems genetics analysis reveals a key regulator of lignin biosynthesis in *Populus deltoides*. *Genome Research* **30**: 1131–1143.

Barakate A, Stephens J, Goldie A, Hunter WN, Marshall D, Hancock RD, Lapierre C, Morreel K, Boerjan W, Halpin C. 2011. Syringyl lignin is unaltered by severe sinapyl alcohol dehydrogenase suppression in tobacco. *Plant Cell* **23**: 4492–4506.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**: 1–48.

Boerjan W, Ralph J, Baucher M. 2003. Lignin biosynthesis. *Annual Review of Plant Biology* **54**: 519–546.

Bolger AM, Lohse M, Usadel B. 2014. TRIMMOMATIC: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

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**: 1153113.

Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L *et al.* 2016. JBROWSE: a dynamic web platform for genome visualization and analysis. *Genome Biology* **17**: 1–12.

Chanoca A, De Vries L, Boerjan W. 2019. Lignin engineering in forest trees. *Frontiers in Plant Science* **10**: 912.

Chantreau M, Tuominen H. 2022. Spatio-temporal regulation of lignification. *Advances in Botanical Research* **104**: 271–316.

Chao N, Liu S-X, Liu B-M, Li N, Jiang X-N, Gai Y. 2014. Molecular cloning and functional analysis of nine cinnamyl alcohol dehydrogenase family members in *Populus tomentosa*. *Planta* **240**: 1097–1112.

Chen C, Meyermans H, Burggraeve B, De Rycke RM, Inoue K, De Vleesschauwer V, Steenackers M, Van Montagu MC, Engler GJ, Boerjan WA. 2000. Cell-specific and conditional expression of caffeoyl-coenzyme A-3-O-methyltransferase in poplar. *Plant Physiology* **123**: 853–867.

Chomczynski P, Sacchi N. 2006. The single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on. *Nature Protocols* **1**: 581–585.

Cooke J, Brown K, Wu R, Davis J. 2003. Gene expression associated with N-induced shifts in resource allocation in poplar. *Plant, Cell & Environment* **26**: 757–770.

Coscia M, Neffke FMH. 2017. Network backboning with noisy data. *2017 IEEE 33rd International Conference on Data Engineering* (*ICDE*): pp. 425–436.

Crivellaro A, Piermattei A, Dolezal J, Dupree P, Büntgen U. 2022. Biogeographic implication of temperature-induced plant cell wall lignification. *Communications Biology* **5**: 767.

De Meester B, Vanholme R, Mota T, Boerjan W. 2022. Lignin engineering in forest trees: from gene discovery to field trials. *Plant Communications* **3**: 100465.

Delhomme N, Mähler N, Schiffthaler B, Sundell D, Mannapperuma C, Hvidsten TR, Street NR. 2014. Guidelines for RNA-Seq data analysis. *Epigenesys Protocol* **67**: 1–24.

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.

Dohm M. 2002. Repeatability estimates do not always set an upper limit to heritability. *Functional Ecology* **16**: 273–280.

Du J, Miura E, Robischon M, Martinez C, Groover A. 2011. The *Populus* Class III HD ZIP transcription factor *POPCORONA* affects cell differentiation during secondary growth of woody stems. *PLoS ONE* **6**: e17458.

Edgar RE. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

Escamez S, Robinson KM, Luomaranta M, Gandla ML, Mähler N, Yassin Z, Grahn T, Scheepers G, Stener L-G, Jansson S *et al.* 2023. Genetic markers and tree properties predicting wood biorefining potential in aspen (Populus tremula) bioenergy feedstock. *Biotechnology for Biofuels and Bioproducts* **16**: 1–16.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MULTIQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.

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

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. 2007. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology* **5**: e8.

Feuillet C, Lauvergeat V, Deswarte C, Pilate G, Boudet A, Grima-Pettenati J. 1995. Tissue- and cell-specific expression of a cinnamyl alcohol dehydrogenase promoter in transgenic poplar plants. *Plant Molecular Biology* **27**: 651–667.

Fornes O, Castro-Mondragon JA, Khan A, Van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D *et al.* 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48**: D87–D92.

Fracheboud Y, Luquez V, Bjorken L, Sjodin A, Tuominen H, Jansson S. 2009. The control of autumn senescence in European aspen. *Plant Physiology* **149**: 1982–1991.

Furches A, Kainer D, Weighill D, Large A, Jones P, Walker AM, Romero J, Gazolla JGFM, Joubert W, Shah M *et al.* 2019. Finding new cell wall regulatory genes in *Populus trichocarpa* using multiple lines of evidence. *Frontiers in Plant Science* **10**: 1249.

Geng P, Zhang S, Liu J, Zhao C, Wu J, Cao Y, Fu C, Han X, He H, Zhao Q. 2020. MYB20, MYB42, MYB43, and MYB85 regulate phenylalanine and lignin biosynthesis during secondary cell wall formation. *Plant Physiology* **182**: 1272–1283.

Gerber L, Eliasson M, Trygg J, Moritz T, Sundberg B. 2012. Multivariate curve resolution provides a high-throughput data processing pipeline for pyrolysis-gas chromatography/mass spectrometry. *Journal of Analytical and Applied Pyrolysis* **95**: 95–100.

Geshi N, Johansen JN, Dilokpimol A, Rolland A, Belcram K, Verger S, Kotake T, Tsumuraya Y, Kaneko S, Tryfona T *et al.* 2013. A galactosyltransferase acting on arabinogalactan protein glycans is essential for embryo development in Arabidopsis. *The Plant Journal* **76**: 128–137.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018.

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.

Guo S, Jiang Q, Chen L, Guo D. 2016. Gene regulatory network inference using PLS-based methods. *BMC Bioinformatics* **17**: 1–10.

Haury A-C, Mordelet F, Vera-Licona P, Vert J-P. 2012. TIGRESS: trustful inference of gene regulation using stability selection. *BMC Systems Biology* **6**: 1–17.

Howles PA, Birch RJ, Collings DA, Gebbie LK, Hurley UA, Hocart CH, Arioli T, Williamson RE. 2006. A mutation in an Arabidopsis ribose 5-phosphate isomerase reduces cellulose synthesis and is rescued by exogenous uridine. *The Plant Journal* **48**: 606–618.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5**: e12776.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**: 275–282.

Kim J, Jung J-H, Reyes JL, Kim Y-S, Kim S-Y, Chung K-S, Kim JA, Lee M, Lee Y, Kim VN *et al.* 2005. microRNA-directed cleavage of ATHB15 mRNA regulates vascular development in Arabidopsis inflorescence stems. *The Plant Journal* **42**: 84–94.

Kolde R. 2015. *Package 'pheatmap'*. R package 1.

Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**: 3211–3217.

Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, Keurentjes JJ, van Eeuwijk FA. 2015. Marker-based estimation of heritability in immortal populations. *Genetics* **199**: 379–398.

Kumar V, Hainaut M, Delhomme N, Mannapperuma C, Immerzeel P, Street NR, Henrissat B, Mellerowicz EJ. 2019. Poplar carbohydrate-active enzymes: whole-genome annotation and functional analyses based on RNA expression data. *The Plant Journal* **99**: 589–609.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 1–13.

Li L, Cheng XF, Leshkevich J, Umezawa T, Harding SA, Chiang VL. 2001. The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. *Plant Cell* **13**: 1567–1586.

Li P, Xiao L, Du Q, Quan M, Song Y, He Y, Huang W, Xie J, Lv C, Wang D. 2023. Genomic insights into selection for heterozygous alleles and woody traits in *Populus tomentosa*. *Plant Biotechnology Journal* **21**: 2002–2018.

Lihavainen J, Šimura J, Bag P, Fataftah N, Robinson KM, Delhomme N, Novák O, Ljung K, Jansson S. 2023. Salicylic acid metabolism and signalling coordinate senescence initiation in aspen in nature. *Nature Communications* **14**: 4288.

Lin S, Miao Y, Huang H, Zhang Y, Huang L, Cao J. 2022. Arabinogalactan proteins: focus on the role in cellulose synthesis and deposition during plant cell wall biogenesis. *International Journal of Molecular Sciences* **23**: 6578.

Liu K-H, Liu M, Lin Z, Wang Z-F, Chen B, Liu C, Guo A, Konishi M, Yanagisawa S, Wagner G *et al.* 2022. NIN-like protein 7 transcription factor is a plant nitrate sensor. *Science* **377**: 1419–1425.

Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2(-\Delta\Delta C(T))$ method. *Methods* **25**: 402–408.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**: 1–21.

Luquez V, Hall D, Albrectsen BR, Karlsson J, Ingvarsson P, Jansson S. 2008. Natural phenological variation in aspen (*Populus tremula*): the SwAsp collection. *Tree Genetics & Genomes* **4**: 279–292.

Mähler N, Schiffthaler B, Robinson KM, Terebieniec BK, Vučak M, Mannapperuma C, Bailey ME, Jansson S, Hvidsten TR, Street NR. 2020. Leaf shape in *Populus tremula* is a complex, omnigenic trait. *Ecology and Evolution* **10**: 11922–11940.

Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. 2017. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genetics* **13**: e1006402.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. 2006. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7**: S7.

Meng X, Pu Y, Yoo CG, Li M, Bali G, Park DY, Gjersing E, Davis MF, Muchero W, Tuskan GA. 2017. An in-depth understanding of biomass recalcitrance using natural poplar variants as the feedstock. *ChemSusChem* **10**: 139–150.

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**: 1–14.

Nica AC, Dermitzakis ET. 2013. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **368**: 20120362.

Nilsson O, Aldén T, Sitbon F, Anthony Little C, Chalupa V, Sandberg G, Olsson O. 1992. Spatial pattern of cauliflower mosaic virus 35S promoter-luciferase expression in transgenic hybrid aspen trees monitored by enzymatic assay and non-destructive imaging. *Transgenic Research* **1**: 209–220.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**: 417–419.

Peterson RA, Cavanaugh JE. 2019. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics* **47**: 2312–2327.

Pilate G, Chabbert B, Cathala B, Yoshinaga A, Leplé J-C, Laurans F, Lapierre C, Ruel K. 2004. Lignification and tension wood. *Comptes Rendus Biologies* **327**: 889–901.

Pitre FE, Pollet B, Lafarguette F, Cooke JE, MacKay JJ, Lapierre C. 2007. Effects of increased nitrogen supply on the lignification of poplar wood. *Journal of Agricultural and Food Chemistry* **55**: 10306–10314.

Porth I, Klapšte J, Skyba O, Hannemann J, McKown AD, Guy RD, DiFazio SP, Muchero W, Ranjan P, Tuskan GA. 2013. Genome-wide association

mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms. *New Phytologist* **200**: 710–726.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Ragauskas AJ, Beckham GT, Biddy MJ, Chandra R, Chen F, Davis MF, Davison BH, Dixon RA, Gilna P, Keller M *et al.* 2014. Lignin valorization: improving lignin processing in the biorefinery. *Science* **344**: 1246843.

Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, Lemma RB, Lucas J, Chèneby J, Baranasic D *et al.* 2024. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **52**: D174–D182.

Renström A, Choudhary S, Gandla ML, Jönsson LJ, Hedenström M, Jämtgård S, Tuominen H. 2024. The effect of nitrogen source and levels on hybrid aspen tree physiology and wood formation. *Physiologia Plantarum* **176**: e14219.

Robinson KM, Schiffthaler B, Liu H, Westman SM, Rendón-Anaya M, Ahlgren Kalman T, Kumar V, Canovi C, Bernhardsson C, Delhomme N *et al.* 2024. An improved chromosome-scale genome assembly and population genetics resource for *Populus tremula*. *bioRxiv*. doi: 10.1101/805614.

Rossi S, Anfodillo T, Menardi R. 2006. Trephor: a new tool for sampling microcores from tree stems. *IAWA Journal* **27**: 89–97.

Rosvall M, Bergstrom CT. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, USA* **105**: 1118–1123.

Ruyssinck J, Huynh-Thu VA, Geurts P, Dhaene T, Demeester P, Saeys Y. 2014. Nimefi: gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS ONE* **9**: e92709.

Schiffthaler B, van Zalen E, Serrano AR, Street NR, Delhomme N. 2023. Seiðr: efficient calculation of robust ensemble gene networks. *Heliyon* **31**: e16811.

Shabalin AA. 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**: 1353–1358.

Shi R, Sun Y-H, Li Q, Heber S, Sederoff R, Chiang VL. 2010. Towards a systems approach for lignin biosynthesis in *Populus trichocarpa*: transcript abundance and specificity of the monolignol biosynthetic genes. *Plant and Cell Physiology* **51**: 144–163.

Song L, Langfelder P, Horvath S. 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* **13**: 1–21.

Stincone A, Prigione A, Cramer T, Wamelink MM, Campbell K, Cheung E, Olin-Sandoval V, Grüning NM, Krüger A, Tauqeer Alam M *et al.* 2015. The return of metabolism: biochemistry and physiology of the pentose phosphate pathway. *Biological Reviews* **90**: 927–963.

Storey J, Bass A, Dabney A, Robinson D. 2020. qvalue: *Q-value estimation for false discovery rate control*. R package v.2.22.0. [WWW document] URL http://github.com/jdstorey/qvalue [accessed 10 October 2020].

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences, USA* **100**: 9440–9445.

Stuart JM, Segal E, Koller D, Kim SK. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**: 249–255.

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.

Sulis DB, Jiang X, Yang C, Marques BM, Matthews ML, Miller Z, Lan K, Cofre-Vega C, Liu B, Sun R. 2023. Multiplex CRISPR editing of wood for sustainable fiber production. *Science* **381**: 216–221.

Sundell D, Mannapperuma C, Netotea S, Delhomme N, Lin Y-C, Sjödin A, Van de Peer Y, Jansson S, Hvidsten TR, Street NR. 2015. The plant genome integrative explorer resource: PlantGenIE.org. *New Phytologist* **208**: 1149–1156.

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.

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

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Research* **40**: e115.

Vanholme R, De Meester B, Ralph J, Boerjan W. 2019. Lignin biosynthesis and its integration into metabolism. *Current Opinion in Biotechnology* **56**: 230–239.

Vanholme R, Morreel K, Darrah C, Oyarce P, Grabber JH, Ralph J, Boerjan W. 2012. Metabolic engineering of novel lignin in biomass crops. *New Phytologist* **196**: 978–1000.

Wang J, Ding J, Tan B, Robinson KM, Michelson IH, Johansson A, Nystedt B, Scofield DG, Nilsson O, Jansson S *et al.* 2018. A major locus controls local adaptation and adaptive life history variation in a perennial plant. *Genome Biology* **19**: 1–17.

Wang JP, Naik PP, Chen H-C, Shi R, Lin C-Y, Liu J, Shuford CM, Li Q, Sun Y-H, Tunlaya-Anukit S. 2014. Complete proteomic-based enzyme reaction and inhibition kinetics reveal how monolignol biosynthetic enzyme families affect metabolic flux and lignin in *Populus trichocarpa*. *Plant Cell* **26**: 894–914.

Wei H, Song Z, Xie Y, Cheng H, Yan H, Sun F, Liu H, Shen J, Li L, He X. 2023. High temperature inhibits vascular development via the PIF4-miR166-HB15 module in Arabidopsis. *Current Biology* **33**: 3203–3214.

Wickham H. 2011. ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics* **3**: 180–185.

Xie M, Muchero W, Bryan AC, Yee K, Guo H-B, Zhang J, Tschaplinski TJ, Singan VR, Lindquist E, Payyavula RS *et al.* 2018. A 5-enolpyruvylshikimate 3-phosphate synthase functions as a transcriptional repressor in *Populus*. *Plant Cell* **30**: 1645–1660.

Xiong Y, DeFraia C, Williams D, Zhang X, Mou Z. 2009. Deficiency in a cytosolic ribose-5-phosphate isomerase causes chloroplast dysfunction, late flowering and premature cell death in Arabidopsis. *Physiologia Plantarum* **137**: 249–263.

Yao T, Zhang J, Yates TB, Shrestha HK, Engle NL, Ployet R, John C, Feng K, Bewg WP, Chen MS *et al.* 2023. Expression quantitative trait loci mapping identified PtrXB38 as a key hub gene in adventitious root development in *Populus*. *New Phytologist* **239**: 2248–2264.

Yoo CG, Yang Y, Pu Y, Meng X, Muchero W, Yee KL, Thompson OA, Rodriguez M, Bali G, Engle NL *et al.* 2017. Insights of biomass recalcitrance in natural *Populus trichocarpa* variants for biomass conversion. *Green Chemistry* **19**: 5467–5478.

Zang L, Zheng T, Chu Y, Ding C, Zhang W, Huang Q, Su X. 2015. Genome-wide analysis of the fasciclin-like Arabinogalactan protein gene family reveals differential expression patterns, localization, and salt stress response in *Populus*. *Frontiers in Plant Science* **6**: 1140.

Zhang J, Tuskan GA, Tschaplinski TJ, Muchero W, Chen JG. 2020. Transcriptional and post-transcriptional regulation of lignin biosynthesis pathway genes in *Populus*. *Frontiers in Plant Science* **11**: 652.

Zhang J, Yang Y, Zheng K, Xie M, Feng K, Jawdy SS, Gunter LE, Ranjan P, Singan VR, Engle N *et al.* 2018. Genome-wide association studies and expression-based quantitative trait loci analyses reveal roles of HCT 2 in caffeoylquinic acid biosynthesis and its regulation by defense-responsive transcription factors in *Populus*. *New Phytologist* **220**: 502–516.

Zhang L, Lu D, Ge X, Du J, Wen S, Xiang X, Du C, Zhou X, Hu J. 2023. Insight into growth and wood properties based on QTL and eQTL mapping in *Populus deltoides* 'Danhong'× *Populus simonii* 'Tongliao1'. *Industrial Crops and Products* **199**: 116731.

Zhang X, Liu K, Liu Z-P, Duval B, Richer J-M, Zhao X-M, Hao J-K, Chen L. 2013. NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics* **29**: 106–113.

Zhao H, Qu C, Zuo Z, Cao L, Zhang S, Xu X, Xu Z, Liu G. 2022. Genome identification and expression profiles in response to nitrogen treatment analysis of the class I CCoAOMT gene family in *Populus*. *Biochemical Genetics* **60**: 656–675.

Zhong R, Allen JD, Xiao G, Xie Y. 2014. Ensemble-based network aggregation improves the accuracy of gene network reconstruction. *PLoS ONE* **9**: e106319.

Zhong R, Lee C, Zhou J, McCarthy RL, Ye Z-H. 2008. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. *Plant Cell* **20**: 2763–2782.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**: 821–824.

Supporting Information

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

Dataset S1 Transcriptomic analysis of SwAsp trees.

Fig. S1 Analysis of population structure in the SwAsp collection.

Fig. S2 A summary of significant gene expression associations after removal of hidden confounders.

Fig. S3 GO terms in lignin clusters.

Fig. S4 The variation in the expression of the lignin-biosynthetic genes in the SwAsp collection.

Fig. S5 Transcription factor binding motifs in the promoter region of the lignin-biosynthetic genes.

Fig. S6 Growth and wood anatomy in *Populus HOMEOBOX PROTEIN HB5* RNAi lines.

Fig. S7 Regulation of *PAL3* by Homeobox protein 5 (HB5).

Fig. S8 Validation of the SwAsp GWAS results in an independent Umeå aspen (UmAsp) collection.

Table S1 Complete Pyrolysis-GC-MS data in the SwAsp and UmAsp collections.

Table S2 Details of the eQTL analysis.

Table S3 Details of the lignin subnetwork.

Table S4 Transcription factors in the lignin subnetwork.

Table S5 The 1000 most significant SNP associations for the lignin-related Pyrolysis-GC-MS traits in the GWAS analysis of SwAsp.

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.