Network analysis reveals the relationship among wood properties, gene expression levels and genotypes of natural *Populus trichocarpa* accessions

Ilga Porth[1,2], Jaroslav Klápště[3,4], Oleksandr Skyba[1], Michael C. Friedmann[2], Jan Hannemann[5], Juergen Ehlting[5], Yousry A. El-Kassaby[3], Shawn D. Mansfield[1] and Carl J. Douglas[2]

[1]Department of Wood Science, University of British Columbia, Vancouver, BC Canada, V6T 1Z4; [2]Department of Botany, University of British Columbia, Vancouver, BC Canada, V6T 1Z4; [3]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC Canada, V6T 1Z4; [4]Department of Dendrology and Forest Tree Breeding, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences, Prague, 165 21, Czech Republic; [5]Department of Biology and Centre for Forest Biology, University of Victoria, Victoria, BC Canada, V8W 3N5

Author for correspondence:
Carl J. Douglas
Tel: +1 604 822 2618
Email: carl.douglas@ubc.ca

Received: *8 March 2013*
Accepted: *17 June 2013*

New Phytologist (2013) **200:** 727–742
doi: 10.1111/nph.12419

Key words: network analysis, phenotype prediction, *Populus trichocarpa*, single nucleotide polymorphisms (SNPs), transcriptomics, wood chemistry, wood density, wood properties.

Summary

- High-throughput approaches have been widely applied to elucidate the genetic underpinnings of industrially important wood properties. Wood traits are polygenic in nature, but gene hierarchies can be assessed to identify the most important gene variants controlling specific traits within complex networks defining the overall wood phenotype. We tested a large set of genetic, genomic, and phenotypic information in an integrative approach to predict wood properties in *Populus trichocarpa*.
- Nine-yr-old natural *P. trichocarpa* trees including accessions with high contrasts in six traits related to wood chemistry and ultrastructure were profiled for gene expression on 49k Nimblegen (Roche NimbleGen Inc., Madison, WI, USA) array elements and for 28 831 polymorphic single nucleotide polymorphisms (SNPs). Pre-selected transcripts and SNPs with high statistical dependence on phenotypic traits were used in Bayesian network learning procedures with a stepwise K2 algorithm to infer phenotype-centric networks.
- Transcripts were pre-selected at a much lower logarithm of Bayes factor (logBF) threshold than SNPs and were not accommodated in the networks. Using persistent variables, we constructed cross-validated networks for variability in wood attributes, which contained four to six variables with 94–100% predictive accuracy.
- Accommodated gene variants revealed the hierarchy in the genetic architecture that underpins substantial phenotypic variability, and represent new tools to support the maximization of response to selection.

Introduction

Wood formation in trees is a complex process under strict developmental control. Its initiation occurs at the vascular cambium by cell differentiation; the stages of secondary xylem (wood) development involve cell division, cell expansion, cell wall thickening by deposition of the polysaccharides cellulose and hemicelluloses and the polyphenolic polymer lignin, and programmed cell death (Plomion *et al.*, 2001). Phytohormones such as auxin, cytokinin, gibberellins and ethylene regulate cambium activity, and xylogenesis is environmentally triggered by photoperiodic and temperature stimuli (Ursache *et al.*, 2012). The regulation of secondary cell wall deposition involves coordinated expression of many genes in a tissue-specific and developmental manner, including genes with regulatory functions (i.e. transcription factors) and structural functions, and this expression is governed by complex gene regulatory networks (Hertzberg *et al.*, 2001; Persson *et al.*, 2005; Demura & Ye, 2010; Ursache *et al.*, 2012; Wang & Dixon, 2012). At the macromolecular level, the relative amount as well as the structure of all three major wood components (cellulose, hemicellulose and lignin) defines the mechanical and biochemical properties of wood that are important to industrial applications. For example, wood density is defined by cell wall thickness, the proportions of thick- to thin-walled cells and the structural organization of cells, which together determine the amount of void volume in the wood (porosity). Also, in angiosperms, the proportion of vessels to fibres has an impact on density, with a higher proportion of fibres to vessels yielding higher density. Lignin content is another important industrial trait, as Kraft pulps derived from wood with high lignin content require extensive bleaching to increase paper quality and durability (Hon, 1981). Wood traits that are of particular interest for the production of cellulosic biofuels such as ethanol are those related to enzymatic release of fermentable sugars following biomass pretreatment (lignin and hemicelluloses content), and yield relative to biomass volume (wood density, relative alpha-cellulose content and relative glucose content).

Populus species (poplars) are fast growing and adapted across broad latitudinal ranges, but have not been domesticated to optimize traits of economic importance. One approach to improving poplars as a feedstock for industrial applications is to use transgenic methods (Vanholme *et al.*, 2010). However, these changes can also drastically impair tree physiology when the extreme phenotype is selected and/or transgenes are constitutively or ectopically expressed (Coleman *et al.*, 2008; Kitin *et al.*, 2010; Voelker *et al.*, 2011). An alternative approach is the use of breeding to capitalize on the substantial variation that exists in wood and lignocellulosic traits in natural and breeding populations. Consequently, there is increasing interest in establishing links between relevant phenotypes and genetic variants (Ingvarsson & Street, 2011), as this information can facilitate rapid genetic improvement of trees as biomass and bioenergy feedstocks via the implementation of marker-assisted breeding (Neale & Kremer, 2011; Harfouche *et al.*, 2012; Nieminen *et al.*, 2012). However, as wood chemical composition and ultrastructure are highly complex traits (Dinus, 2000), dissection of the genetic architecture underlying such variation in cell wall traits is challenging, and the current knowledge of genetic underpinnings that underlie essential biofuel and biomass traits is incomplete.

Gene expression profiling studies using microarrays have been widely used to examine the process of wood formation in poplar (e.g., Hertzberg *et al.*, 2001; Schrader *et al.*, 2004) and in *Pinus radiata*, gene expression analysis of individuals with contrasting wood properties identified differentially expressed candidate genes that may directly relate transcriptional regulation to particular wood traits (Li *et al.*, 2011, 2012). Such approaches offer the opportunity to recover a distinct transcriptome correlated to the trait under study (Li *et al.*, 2011, 2012).

In the search for candidate genes whose allelic variants may impact wood formation and wood traits, quantitative genetics approaches have been pursued. These approaches encompass quantitative trait locus (QTL) studies which dissect phenotypic trait variation in controlled crosses by linkage to segregating genetic markers (QTLs; Pot *et al.*, 2002; Kirst *et al.*, 2004; Novaes *et al.*, 2009; Thumma *et al.*, 2010), and association genetics studies that identify significant associations between trait variation and single nucleotide polymorphisms (SNPs) in ‘unstructured’ populations (Gonzalez-Martinez *et al.*, 2007; Thumma *et al.*, 2009; Wegrzyn *et al.*, 2010; Beaulieu *et al.*, 2011; Porth *et al.*, 2013a). However, these approaches cannot be easily used to elucidate how genes and gene variants interact with each other, nor can they achieve the goal of effectively predicting combined effects of genetic variants on a phenotype. Gene expression network analysis provides a method for elucidating such gene–gene interactions (Bassel *et al.*, 2012). Starting from approaches that determine interactions based on Pearson correlations between gene-pair expression levels to describe co-regulation (relevance networks), partial correlations have been computed to assess *direct* interactions between gene pairs conditioned on the behaviour of other genes (Ma *et al.*, 2007).

In the present study, we undertook a different approach, that is, to infer for each studied wood trait a network of genetic variables that conjointly represent the most accurate predictor for the substantial natural variation in a particular attribute. Thus, we inferred Bayesian networks of genetic variants (SNPs) and expression variation with all nodes linked to the phenotype, that is, directed and phenotype-centric networks (Chang & McGeachie, 2011), in our study population. By testing each transcript and each SNP individually for their likelihood of dependence on the phenotype, we introduced only highly informative variables into the network learning approach. We investigated a set of black cottonwood (*Populus trichocarpa*) accessions for transcriptome differences using a 49k Nimblegen (Roche NimbleGen Inc., Madison, WI, USA) poplar gene expression microarray. These accessions were selected on the basis of highly contrasting wood properties (Porth *et al.*, 2013b) that were grouped according to their phenotypic properties, allowing pair-wise comparisons of individuals with contrasting extremes in wood phenotypes. We identified transcripts that showed statistically significant differential expression levels in the developing xylem of individuals with contrasting wood properties and then used the genotyping information from a set of 28 831 polymorphic SNPs in 3543 candidate genes (Geraldes *et al.*, 2013; Porth *et al.*, 2013a) as well as the expression levels of genes represented on the 49K Nimblegen array to test whether a variable (continuous transcript or discrete SNP) showed statistical dependence on a wood attribute. The identified variables were then used in a network learning approach and phenotype-centric networks were computed according to the approach of Chang & McGeachie (2011). Phenotype-centric networks were validated by predicting each phenotype based on persistent variables included in the networks, demonstrating the utility of this approach in predicting relationship between genotypes and wood properties in natural *P. trichocarpa* accessions.

Materials and Methods

Black cottonwood samples

Black cottonwood (*Populus trichocarpa* Torr. & Gray) individuals were grown in a common garden in Surrey, BC, Canada. The origin of the samples is described in Xie *et al.* (2009) and their geographical origin is displayed in Fig. 1. In 2008, wood cores were taken from 9-yr-old trees and phenotyped for wood chemistry and ultrastructure traits (Porth *et al.*, 2013b). Details of the wood trait analysis are found in Porth *et al.* (2013b). The individuals selected for the present study were identified as trees ranking very high and very low, respectively, in a range-wide distribution of phenotypic values in *P. trichocarpa.* The range of phenotypic values in these individuals can be found in Supporting Information Table S1.

SNP genotyping

Genotyping of *P. trichocarpa* individuals was carried out using the Illumina (Illumina Inc., San Diego, CA, USA) 34k *Populus* SNP genotyping array with 34 131 SNPs from 3543 candidate genes, described in detail by Geraldes *et al.* (2013). We eliminated SNPs

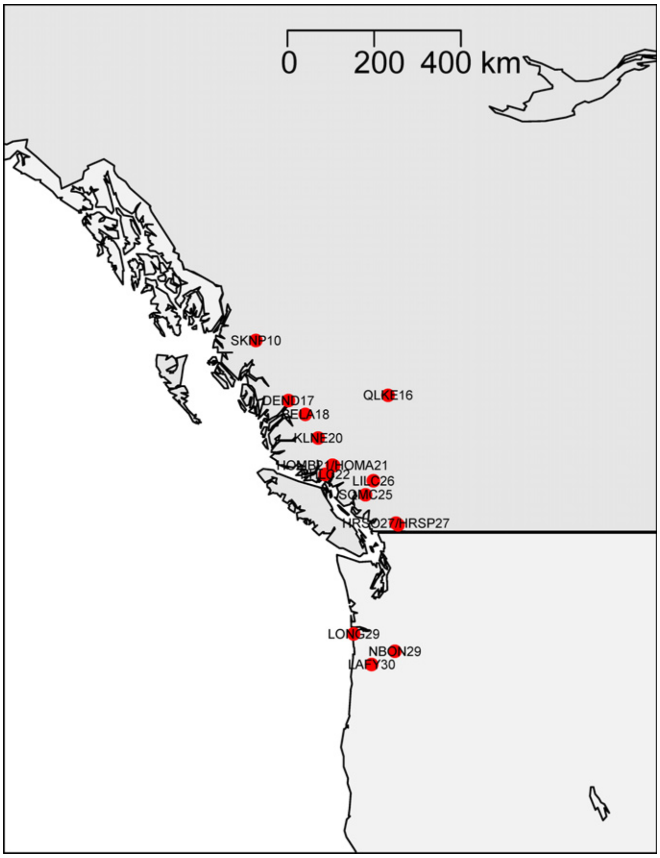

Fig. 1 Geographical origins of 16 *Populus trichocarpa* accessions with two biological replicates in the common-garden field (Surrey, BC, Canada) used for microarray expression profiling and subsequent analyses. Displayed are the 15 different drainages located in the Pacific Northwest (BC, Canada; Oregon, USA) from which the samples originated; note that drainage LONG29 has two samples; for details about the classifications of each sample in terms of the five wood chemistry traits and one ultrastructure trait, see Supporting Information Tables S1 and S2.

with: minor allele frequency below 0.05; > 10% missing values; and Illumina's GeneTrain score below 0.5. These selection criteria reduced the total number of SNPs genotyped to 29 233.

Source of RNA for expression profiling

We identified poplar individuals with the most extreme phenotypes in six wood traits (relative glucose, alpha-cellulose, xylose, total hemicellulose and lignin contents, and average wood density) using phenotype data (Porth *et al.*, 2013b) from 384 9-yr-old *P. trichocarpa* trees grown in a common-garden plot (located 49.18°N and 122.85°W; Surrey, BC, Canada) representing populations that span a latitudinal range of 44.0–58.6°N. In total, 17 genotypes with two biological replicates, and two additional genotypes without a second replicate, with contrasting wood phenotypes were individually profiled for transcriptome analysis (Table S1); the experimental set-up is shown in Table S2. Developing xylem was harvested at breast height from the north side of each tree stem on the morning of 9 June 2011 by first removing a window of bark that also contained cambium cells from the stem to expose the underlying current-year developing xylem. This material, but not underlying mature xylem, was scraped with uniform pressure using sterile razor blades to remove several layers of developing, but not mature, secondary xylem cells. The tissue was immediately flash-frozen in liquid nitrogen and stored at −80°C until used. RNA extractions and quality control of the extracted RNA were carried out as described previously (Geraldes *et al.*, 2011). Total RNA quality was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA, USA) before microarray analysis. Samples with an RNA integrity number (RIN) value of ≥ 8.0 were used for microarray analysis.

Gene expression profiling and analysis

For gene expression analysis, we employed the Populus Gene Expression 12x135k Array (#GPL16717) (Roche NimbleGen) based on 48 146 nuclear gene models from version 1.1 of the *Populus trichocarpa* genome assembly and annotation (described in Tsai *et al.*, 2011). We re-annotated array features according to version 2.2 of the *Populus trichocarpa* genome assembly and annotation, which includes 34 845 unique version 2.2 gene models and additional annotations based on similarity to *Eucalyptus grandis* genes (Table S3). Microarray hybridizations and basic bioinformatics analysis were carried out at the Nimblegen processing site (Jack Bell Research Centre, Prostate Centre, Vancouver, BC, Canada). RNA samples were hybridized to three NimbleGen 12-plex poplar arrays totalling 36 hybridizations. Samples were prepared following NimbleGen's Arrays User Guide Gene Expression Analysis version 3.2. In brief, 10 μg of each total RNA was converted to double-stranded (ds) cDNA with the Invitrogen SuperScript Double-Stranded cDNA Synthesis Kit, 1 μg of each ds cDNA was fluorescently labelled using the NimbleGen One-Color DNA Labeling Kit, and 4 μg of each Cy3-labelled sample was hybridized on Roche Nimblegen poplar gene expression microarrays (Design Name Populus 135K EXP HX12 090828). Arrays were scanned at a 5 μm resolution with the Molecular Devices (Molecular Devices LLC, Sunnyvale, CA, USA) GenePix 4200AL scanner. NIMBLESCAN version 2.5 (Roche NimbleGen) was used for quantitation and RMA normalization of data, which included quantile normalization and background subtraction. Agilent's GENESPRING 7.3.1 (Agilent Technologies Inc., Santa Clara, CA, USA) was used to analyze the normalized data. The microarray signal intensities were deposited at National Center for Biotechnology (NCBI) GEO database #GSE44606.

To find significantly up- or down-regulated genes, fold changes between the compared groups and *P*-values obtained from *t*-tests comparing the same groups were calculated. The *t*-tests were performed on normalized data that had been log-transformed and the variances were not assumed to be equal between sample groups (Welch test). We applied a relatively low *P*-value ($P < 0.05$) and a lower fold-change cut-off ($FC > 1.2$) than many microarray-based expression profiling studies, similar to Li *et al.* (2011), because small variations in the expression of many genes were expected in the study population. Biological replicates (including the ramets) were treated independently; that is, they were not averaged in the *t*-test analysis.

Random Forest, a classification algorithm (Breiman, 2001), was recently introduced for pre-selecting genes from microarray experiments (Diaz-Uriarte & de Andres, 2006). This method was shown to be robust to the inherent characteristics of microarray data (noisy expression phenotypes; number of variables largely exceeds number of observations), and provides the advantage of being applicable to multi-class problems for class prediction. We tested predictabilities using this algorithm optimized for microarray data (Diaz-Uriarte & de Andres, 2006) by running R scripts implemented in the web version http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html. SNPs were pre-selected using R scripts (Bureau *et al.*, 2005).

Integrative network analysis

A set of 16 different genotypes (each with two ramets) with available genotype information were assigned to high, medium, and low phenotype subclasses, and expression levels of duplicate genotypes in each class were averaged using both biological replicates (Table S4). A Bayesian network-based analysis was performed on gene expression, SNP and wood phenotype data using scripts in MATLAB (MathWorks Inc., Natick, MA, USA) kindly provided by Drs H. H. Chang and M. McGeachie (Harvard Medical School, MA, USA). We used expression data from *c.* 49k Nimblegen array elements and genotype information involving 28 831 polymorphic SNPs. Missing genotype values were replaced by imputed values using the R-package SYNBREED (Wimmer *et al.*, 2012). The network analysis followed the methodologies previously described (Ferrazzi *et al.*, 2007; Chang & McGeachie, 2011). Briefly, we recoded genotypes as homozygous for major allele frequency (0), heterozygous (1), and homozygous for minor allele frequency (2). First, pre-selection of transcripts and SNPs identified variables with a high probability of being statistically dependent on the phenotypic trait. We performed Gaussian linear regression using R scripts within the library MCMCpack (Martin & Quinn, 2007) to determine the log Bayes factor that provides information about the model fit using the tested variable (in model M_2) against the intercept model with solely the population mean in the regression equation (i.e. reduced model M_1). Hence, the Bayes factor was determined as follows:

$$\mathrm{BF} = \frac{p(y|M_2)}{p(y|M_1)}$$

Then, the selected variables were used in a Bayesian network learning procedure (stepwise K2 algorithm) to infer a phenotype-centric network that best explains the pre-selected SNPs and gene expression levels previously conditioned on the quantitative trait. An important convention is the directionality of the flow of genetic information (DNA → RNA) and setting the maximum number of parents for each node. In sum, this integrative network analysis has the potential to uncover the following interactions: SNP–SNP, SNP–transcript, transcript–transcript, phenotype–SNP, and phenotype–transcript. The individual variables directly associated with the phenotype in the network were used in a local propagation algorithm to test the predictive accuracy of the network (Chang & McGeachie, 2011). Networks were displayed using YED GRAPH EDITOR software (YED 3.9.2; yWorks 2012, yWorks GmbH, Tuebingen, Germany).

Results

Differences in wood attributes are associated with highly complex transcriptome variation

Using the 49k Nimblegen microarray, we profiled gene expression in the developing xylem of *P. trichocarpa* (poplar) samples that represent independent poplar genotypes with extreme phenotypes. The trees grown in a common garden originated from individual drainages in southwestern British Columbia and northwestern USA (Fig. 1). We collected two ramets from most genotypes for RNA isolation, compared the gene expression profiles of individuals that grouped together according to their phenotypic properties, and performed pair-wise comparisons between groups of individuals with highly contrasting wood attributes (see Table S2 for microarray hybridization experimental design). First, we applied a univariate ranking method to prioritize transcripts for further investigation. To identify differentially transcribed genes in developing xylem of individuals with contrasting wood properties, fold-change values between the compared groups and *P*-values derived from *t*-tests comparing the same groups were calculated. The *t*-tests were performed on log-transformed normalized data. Table 1 summarizes significant differences in developing xylem gene expression between trees in contrasting trait bins (high versus low for five trait comparisons). Comprehensive data from all performed *t*-tests can be found in Table S5, including significant increases or decreases in steady-state transcript abundance that indicate mode of gene action. These data revealed extensive gene expression differences in steady-state transcript abundance between poplar individuals with contrasting wood attributes, and suggest that gene expression variation could underlie some of the phenotypic variation observed between these individuals. Results using the false discovery rate (FDR) approach for multiple testing corrections across all 49k genes from the array are summarized in Table S6 (a). In contrast to univariate methods which provide low power for variables with marginal effects in the population, but large interaction effects, Random Forest, a clustering method in phenotype class prediction, is capable of detecting higher order interactions among variables (Lunetta *et al.*, 2004). Thus, we tested the Random Forest algorithm for gene selection (Diaz-Uriarte & de Andres, 2006), and transcripts relevant in class prediction were identified (Table 1) and are summarized in Table S6(b). For each phenotype the prediction error rate (0.632+ bootstrap method; Efron & Tibshirani, 1997) was estimated: alpha-cellulose (0.39295), glucose (0.35722), density (0.46731), lignin (0.38893), hemicelluloses (0.441449), and xylose (0.44296). Finally, a third approach involving regression on phenotypes was performed to identify relevant transcripts and is detailed in the following section.

Table 1 Comparison of different methods for variable (genes and single nucleotide polymorphisms (SNPs)) pre-selection conditional on phenotype variation in *Populus trichocarpa*

Trait	Variable pre-selection type	No. of genes[1]	No. of SNPs
Alpha-cellulose	*t*-test[*,2]	1163 (6)	na
	RF[3]	9	316
	logBF[4,5]	5	72
Glucose	*t*-test[*,2]	984 (2)	na
	RF[3]	40	413
	logBF[4,5]	7	73
Density	*t*-test[*,2]	823 (1)	na
	RF[3]	6	313
	logBF[4,5]	1	78
Lignin	*t*-test[*,2]	1151 (6)	na
	RF[3]	7	432
	logBF[4,5]	3	141
Hemicellulose	*t*-test[*,2]	1781 (22)	na
	RF[3]	32	303
	logBF[4,5]	8	49
Xylose	*t*-test[*,2]	1781 (22)	na
	RF[3]	120	376
	logBF[4,5]	10	56

FC, fold change; FDR, false discovery rate; na, not applicable; RF, Random Forest; logBF, logarithm of Bayes factor.

*P-value < 0.05, FC > 1.2; in brackets is the number of genes at $q < 0.1$; FDR estimation following Storey & Tibshirani (2003); the number of trees/arrays used in each comparison can be found in Supporting Information Tables S2 and S4; for *t*-tests, the number of unigenes is shown; note that for the two class comparisons, that is, high/low in *t*-tests, for hemicelluloses and xylose phenotypes the same individuals were used.

[1] Overlap of variables (genes) among methods provided in Fig. S1.

[2] Li *et al.* (2011, 2012).

[3] Diaz-Uriarte & de Andres (2006) and Breiman (2001).

[4] logBF > 0 (genes), logBF > 2 (SNPs).

[5] Chang & McGeachie (2011).

Identification of transcripts and SNPs statistically dependent on variation in the wood attributes

To extend the above observations, we attempted to identify combinations of transcript abundance from microarray expression profiling using the 49k Nimblegen poplar microarray and SNPs that were related to a given wood attribute. For this analysis, 16 genotypes (accessions) were assigned to high, medium, and low phenotype subclasses for each trait. Genotypes of each accession were determined for a set of 28 831 polymorphic high-quality SNPs derived from hybridization to a 34k Illumina *Populus* genotyping array (see the Materials and Methods section; Geraldes *et al.*, 2013; Porth *et al.*, 2013a). However, to avoid over-fitting of the model in the subsequent Bayesian network analysis, we decided to perform log Bayes factor (logBF) variable pre-selection for both transcripts as well as SNPs. The logBF threshold criteria used to identify variables (transcript abundance and SNPs) that were highly dependent on phenotype were logBF > 0 for transcripts and logBF > 2 for SNPs. Different BF criteria for SNPs and transcripts were chosen according to Chang & McGeachie (2011).

This approach identified several SNPs and transcripts related to phenotypic variation in the traits measured; there were 78 SNPs (located within 51 different genes) and one transcript for wood density, 72 SNPs (within 52 genes) and five transcripts for relative alpha-cellulose content, 73 SNPs (within 59 genes) and seven transcripts for relative glucose content, 49 SNPs (within 36 genes) and eight transcripts for relative hemicellulose content, 56 SNPs (within 42 genes) and 10 transcripts for relative xylose content, and finally 141 SNPs (within 112 genes) and three transcripts for relative lignin content. Tables S6(c) and S7 show the pre-selected transcripts and all pre-selected SNPs, respectively, while Table 2 summarizes the most significant SNPs with logBF > 4 along with all pre-selected transcripts for each phenotype. All identified transcripts for phenotypic variation in wood density, alpha-cellulose/glucose content, and lignin content were significant in the *t*-tests ($P < 0.05$) for differential expression (see section above), and were highly ranked with respect to *P*-values for differential expression. However, for hemicellulose content, we identified an additional transcript not identified as significant for differential expression, and for xylose, we identified four additional transcripts not identified as significant for differential expression (no *t*-test for xylose was performed individually) (Fig. S1). Using more stringent selection criteria in the *t*-tests (q-value < 0.1), we identified a set of 23 transcripts that were validated across different pre-selection analyses (Table 3), with POPTR_0006s28070 (unknown protein) validated across all three analyses and associated with lignin, alpha-cellulose, and glucose variation in both *t*-tests and RF cluster analyses, as well as with lignin and alpha-cellulose in the logBF regression analyses. These individual transcripts identified in our study can potentially be important new candidates for follow-up studies.

In general, the identified transcripts (Table 2) were encoded by genes with functions not previously connected to secondary cell wall formation. By contrast, a large number of the identified SNPs are located within genes implicated in wood and secondary cell wall formation (Table S8), which may be a consequence of the bias towards candidate genes in these processes in the design of the genotyping array (Geraldes *et al.*, 2013). In fact, only two of the poplar genes identified through transcriptome analyses (log BF analyses; Table 2) were represented on the 34k SNP genotyping array (Geraldes *et al.*, 2013).

Among the genes with SNPs correlated to strong phenotypic variation we identified (logBF criteria), 95 had been included on the 34k SNP *Populus* genotyping array based on expression patterns or potential involvement with secondary cell wall formation. The portion of these secondary wall development candidate genes within each set of identified SNPs was on average 33% of the total, but ranged between 30% (glucose content) and 38% (xylose content) for individual traits. A total of 391 SNPs associated with 274 different poplar gene models were identified by pre-selection, and 17% of the pre-selected SNPs were identified for more than one trait. One explanation for this is the high correlation among traits for which we identified individuals at phenotypic extremes. For example, the phenotypic correlation (r) was −0.87 between alpha-cellulose and lignin contents, and almost half of the identified SNPs for alpha-cellulose content were identical to SNPs identified for lignin content. A high

Table 2 Most significant single nucleotide polymorphism (SNP) variables (logarithm of Bayes factor (logBF) > 4) and transcript expression identified by pre-selection as dependent on phenotype in *Populus trichocarpa*

Trait	Transcripts[1]/SNP symbol[2]	Gene model[3]	SNP feature	Coding direction (v2.2)	Amino acids	Coding effect	POPGENIE BestAraHit	POPGENIE description
Density	**fgenesh1_pg.C_scaffold_28000006**	na	na	na	na	na	na	na
Density	scaffold_1_12114727	POPTR_0001s15280	NC	–	–	–	AT1G02130	ARABIDOPSIS RAS 5 (ARA-5); GTP binding
Density	scaffold_1_12437833	POPTR_0001s15580	INT	+	–	–	AT5G47430	Zinc ion binding
Density	scaffold_1_12439669	POPTR_0001s15580	INT	+	–	–	AT5G47430	Zinc ion binding
Density	scaffold_1_12442384	POPTR_0001s15580	INT	+	–	–	AT5G47430	Zinc ion binding
Density	scaffold_1_26389194	POPTR_0001s27480	INT	–	–	–	AT3G08580	ADP/ATP CARRIER 1 (AAC1); ATP:ADP antiporter/binding
Density	scaffold_1_26390177	POPTR_0001s27480	INT	–	–	–	AT3G08580	ADP/ATP CARRIER 1 (AAC1); ATP:ADP antiporter/binding
Density	scaffold_1_47245120	POPTR_0001s46870	NC	–	–	–	AT5G62690	TUB2; GTP binding/GTPase/structural molecule
Density	scaffold_2_8015702	POPTR_0002s10990	INT	–	–	–	AT5G26780	SERINE HYDROXYMETHYLTRANSFERASE 2 (SHM2)
Density	scaffold_6_9686849	POPTR_0006s12530	NC	+	–	–	AT5G57620	myb domain protein 36 (MYB36); DNA binding/transcription factor
Density	scaffold_7_13868213	POPTR_0007s13860	INT	+	–	–	AT3G51150	Kinesin motor family protein
Density	scaffold_7_13868474	POPTR_0007s13860	INT	+	–	–	AT3G51150	Kinesin motor family protein
Density	scaffold_7_13870208	POPTR_0007s13860	INT	+	–	–	AT3G51150	Kinesin motor family protein
Density	scaffold_8_2306887	POPTR_0008s04140	CDS	–	N/N	Syn	AT3G07990	Serine carboxypeptidase-like 27 (SCPL27)
Density	scaffold_9_6023636	POPTR_0009s06150	NC	+	–	–	AT5G13780	GCN5-related N-acetyltransferase, putative
Density	scaffold_12_1814164	POPTR_0012s02170	INT	+	–	–	AT3G49220	Pectinesterase family protein
Density	scaffold_12_1814218	POPTR_0012s02170	INT	+	–	–	AT3G49220	Pectinesterase family protein
Density	scaffold_13_10511534	POPTR_0013s10370	NC	–	–	–	AT4G24340	Phosphorylase family protein
Density	scaffold_18_12087033	POPTR_0018s11280	INT	+	–	–	AT4G30710	Unknown protein
Alpha-cellulose	**POPTR_0006s28070.1**	POPTR_0006s28070	na	na	na	na	AT4G30630	Unknown protein
Alpha-cellulose	**POPTR_0010s17910.1**	POPTR_0010s17910	na	na	na	na	No arabidopsis blast hit	na
Alpha-cellulose	**POPTR_0015s15300.1**	POPTR_0015s15300	na	na	na	na	AT1G16470	Proteasome subunit PAB1
Alpha-cellulose	**POPTR_0006s16720.1**	POPTR_0006s16720	na	na	na	na	AT4G23160	Cysteine-rich RLK (RECEPTOR-like protein kinase) 8
Alpha-cellulose	**POPTR_0004s07230.1**	POPTR_0004s07230	na	na	na	na	AT4G25700	Beta-hydroxylase 1
Alpha-cellulose	scaffold_1_15694837	POPTR_0001s18640	INT	–	–	–	AT1G52190	Proton-dependent oligopeptide transport (POT) family protein
Alpha-cellulose	scaffold_2_358872	POPTR_0002s00780	NC	–	–	–	AT1G76410	ATL8; protein binding/zinc ion binding
Alpha-cellulose	scaffold_2_1322346	POPTR_0002s02340	INT	+	–	–	AT1G20010	TUB5; structural constituent of cytoskeleton
Alpha-cellulose	scaffold_5_23631116	POPTR_0005s25350	CDS	+	V/F	Non-Syn	AT1G75430	BEL1-LIKE HOMEODOMAIN 11 (BLH11); transcription factor
Alpha-cellulose	scaffold_5_23631273	POPTR_0005s25350	INT	+	–	–	AT1G75430	BEL1-LIKE HOMEODOMAIN 11 (BLH11); transcription factor
Alpha-cellulose	scaffold_10_12078425	POPTR_0010s11950	NC	+	–	–	AT1G14890	Enzyme inhibitor/pectinesterase/pectinesterase inhibitor
Glucose	**POPTR_0013s04650.1**	POPTR_0013s04650	na	na	na	na	AT4G13870	Werner syndrome-like exonuclease
Glucose	**eugene3.04220002**	na	na	na	na	na	na	na
Glucose	**POPTR_0010s17910.1**	POPTR_0010s17910	na	na	na	na	No arabidopsis blast hit	na
Glucose	**POPTR_0015s06200.1**	POPTR_0015s06200	na	na	na	na	No arabidopsis blast hit	na
Glucose	**POPTR_0003s15790.1**	POPTR_0003s15790	na	na	na	na	AT3G16560	Protein phosphatase 2C family protein
Glucose	**POPTR_0271s00210.1**	POPTR_0271s00210	na	na	na	na	AT3G17080	Plant self-incompatibility protein S1 family
Glucose	**POPTR_0004s05440.1**	POPTR_0004s05440	na	na	na	na	AT2G31130	Expressed protein
Glucose	scaffold_1_31121055	POPTR_0001s32810	NC	+	–	–	AT4G13980	AT-HSFA5; DNA binding/transcription factor
Glucose	scaffold_1_44690812	POPTR_0001s44250	NC	–	–	–	AT5G54690	GALACTURONOSYLTRANSFERASE 12 (GAUT12)

Table 2 (Continued)

Trait	Transcripts[1]/SNP symbol[2]	Gene model[3]	SNP feature	Coding direction (v2.2)	Amino acids	Coding effect	POPGENIE BestAraHit	POPGENIE description
Glucose	scaffold_2_358872	POPTR_0002s00780	NC	–	–	–	AT1G76410	ATL8; protein binding/zinc ion binding
Glucose	scaffold_5_3509460	NA	CDS	+	T/T	Syn	NA	
Glucose	scaffold_5_23732234	POPTR_0005s25490	NC	–	–	–	AT3G54140	PEPTIDE TRANSPORTER 1 (PTR1)
Glucose	scaffold_6_5969767	POPTR_0006s08210	INT	–	–	–	AT3G53670	Unknown protein
Glucose	scaffold_10_17837178	POPTR_0010s20100	INT	+	–	–	AT3G55770	LIM domain-containing protein
Hemicellulose	**POPTR_0008s15860.1**	POPTR_0008s15860	na	na	na	na	AT5G43810	Stabilizer of iron transporter SufD/Polynucleotidyl transferase
Hemicellulose	**POPTR_0018s12580.1**	POPTR_0018s12580	na	na	na	na	AT5G35200	ENTH/ANTH/VHS superfamily protein
Hemicellulose	**POPTR_0004s18330.1**	POPTR_0004s18330	na	na	na	na	No arabidopsis blast hit	na
Hemicellulose	**POPTR_0004s07020.1**	POPTR_0004s07020	na	na	na	na	No arabidopsis blast hit	na
Hemicellulose	**POPTR_0015s10460.1**	POPTR_0015s10460	na	na	na	na	AT5G20090	Uncharacterised protein family (UPF0041)
Hemicellulose	**eugene3.04400002**	na	na	na	na	na	na	na
Hemicellulose	**POPTR_0001s13400.1**	POPTR_0001s13400	na	na	na	na	AT5G43280	delta(3,5),delta(2,4)-dienoyl-CoA isomerase 1
Hemicellulose	**POPTR_0002s05980.1**	POPTR_0002s05980	na	na	na	na	AT1G06500	Unknown protein
Hemicellulose	scaffold_1_8215615	POPTR_0001s10560	INT	–	–	–	AT1G64390	*Arabidopsis thaliana* glycosyl hydrolase 9C2 (AtGH9C2)
Hemicellulose	scaffold_1_43171555	POPTR_0001s43080	3′-UTR	+	–	–	AT4G20430	Subtilase family protein
Hemicellulose	scaffold_2_7294482	POPTR_0002s10150	3′-UTR	+	–	–	AT1G72230	Plastocyanin-like domain-containing protein
Hemicellulose	scaffold_4_5768592	POPTR_0004s06940	NC	–	–	–	AT4G23570	SGT1A; protein binding
Hemicellulose	scaffold_4_18598531	POPTR_0004s18860	NC	+	–	–	AT2G17040	Arabidopsis NAC domain containing protein 36 (ANAC036); transcription factor
Hemicellulose	scaffold_5_4521110	POPTR_0005s06740	INT	+	–	–	AT2G13440	Glucose-inhibited division family A protein
Hemicellulose	scaffold_5_8660164	POPTR_0005s11880	CDS	–	I/I	Syn	AT3G51150	Kinesin motor family protein
Hemicellulose	scaffold_5_22901037	POPTR_0005s24300	INT	–	–	–	AT3G58170	BET1P/SFT1P-LIKE PROTEIN 14A (BS14A); SNAP receptor/ protein transporter
Hemicellulose	scaffold_13_259799	POPTR_0013s00550	NC	–	–	–	AT3G17390	METHIONINE OVER-ACCUMULATOR 3 (MTO3); methionine adenosyltransferase
Hemicellulose	scaffold_13_4210414	POPTR_0013s05850	CDS	–	I/V	Non-Syn	AT5G17920	ATMS1; 5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase/methionine synthase
Hemicellulose	scaffold_16_5719734	POPTR_0016s07890	CDS	+	Q/Q	Syn	AT5G22400	rac GTPase activating protein, putative
Hemicellulose	scaffold_18_3035218	POPTR_0018s03450	NC	+	–	–	AT4G31550	WRKY11; calmodulin binding/transcription factor
Hemicellulose	scaffold_18_5353831	POPTR_0018s05770	NC	–	–	–	AT3G24520	AT-HSFC1; DNA binding/transcription factor
Hemicellulose	scaffold_19_6711559	POPTR_0019s06140	NC	+	–	–	AT3G45400	Exostosin family protein
Hemicellulose	scaffold_19_6711617	POPTR_0019s06140	NC	+	–	–	AT3G45400	Exostosin family protein
Hemicellulose	scaffold_19_9829010	POPTR_0019s08330	INT	+	–	–	AT5G17010	Sugar transporter family protein
Xylose	**POPTR_0017s01420.1**	POPTR_0017s01420	na	na	na	na	AT3G14460	LRR and NB-ARC domains-containing disease resistance protein
Xylose	**POPTR_0009s00200.1**	POPTR_0009s00200	na	na	na	na	No arabidopsis blast hit	na
Xylose	**POPTR_0002s22560.1**	POPTR_0002s22560	na	na	na	na	AT2G30950	FtsH extracellular protease family
Xylose	**POPTR_0018s06350.1**	POPTR_0018s06350	na	na	na	na	AT5G19950	Domain of unknown function (DUF1767)
Xylose	**POPTR_0015s15100.1**	POPTR_0015s15100	na	na	na	na	AT4G08850	Leucine-rich repeat receptor-like protein kinase family protein
Xylose	**POPTR_0008s15860.1**	POPTR_0008s15860	na	na	na	na	AT5G43810	Stabilizer of iron transporter SufD/polynucleotidyl transferase
Xylose	**POPTR_0007s02810.1**	POPTR_0007s02810	na	na	na	na	AT4G16740	Terpene synthase 03

Table 2 (Continued)

Trait	Transcripts[1]/SNP symbol[2]	Gene model[3]	SNP feature	Coding direction (v2.2)	Amino acids	Coding effect	POPGENIE BestAraHit	POPGENIE description
Xylose	**POPTR_0004s07020.1**	POPTR_0004s07020	na	na	na	na	No arabidopsis blast hit	na
Xylose	**POPTR_0018s13150.1**	POPTR_0018s13150	na	na	na	na	AT2G41480	Peroxidase superfamily protein
Xylose	**fgenesh1_pg.C_LG_VII000651**	na	na	na	na	na	na	na
Xylose	scaffold_1_43171555	POPTR_0001s43080	3′-UTR	+	–	–	AT4G20430	Subtilase family protein
Xylose	scaffold_4_18598531	POPTR_0004s18860	NC	+	–	–	AT2G17040	Arabidopsis NAC domain containing protein 36 (anac036); transcription factor
Xylose	scaffold_5_4521110	POPTR_0005s06740	INT	+	–	–	AT2G13440	Glucose-inhibited division family A protein
Xylose	scaffold_13_4210414	POPTR_0013s05850	CDS	–	I/V	Non-Syn	AT5G17920	ATMS1; 5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase/methionine synthase
Xylose	scaffold_19_6711559	POPTR_0019s06140	NC	+	–	–	AT3G45400	Exostosin family protein
Xylose	scaffold_19_6711617	POPTR_0019s06140	NC	+	–	–	AT3G45400	Exostosin family protein
Xylose	scaffold_19_6945893	POPTR_0019s06290	NC	–	–	–	AT1G08810	myb domain protein 60 (MYB60); DNA binding/transcription factor
Lignin	**POPTR_0015s15300.1**	POPTR_0015s15300	na	na	na	na	AT1G16470	Proteasome subunit PAB1
Lignin	**gw1.IV.50.1**	na	na	na	na	na	na	na
Lignin	**POPTR_0006s28070.1**	POPTR_0006s28070	na	na	na	na	AT4G30630	Unknown protein
Lignin	scaffold_1_2936883	POPTR_0001s03670	NC	–	–	–	AT1G15950	CINNAMOYL COA REDUCTASE 1 (CCR1)
Lignin	scaffold_1_44689823	POPTR_0001s44250	INT	–	–	–	AT5G54690	GALACTURONOSYLTRANSFERASE 12 (GAUT12)
Lignin	scaffold_2_1322346	POPTR_0002s02340	INT	+	–	–	AT1G20010	TUB5; structural constituent of cytoskeleton
Lignin	scaffold_2_10470798	POPTR_0002s14120	INT	–	–	–	AT2G44670	Senescence-associated protein-related
Lignin	scaffold_3_16181354	POPTR_0003s16610	INT	+	–	–	AT5G13120	Peptidyl-prolyl cis-trans isomerase cyclophilin-type family protein
Lignin	scaffold_5_1720220	POPTR_0005s02700	INT	–	–	–	AT3G05420	ACYL-COA BINDING PROTEIN 4 (ACBP4)
Lignin	scaffold_5_1723773	POPTR_0005s02700	INT	–	–	–	AT3G05420	ACYL-COA BINDING PROTEIN 4 (ACBP4)
Lignin	scaffold_5_1724721	POPTR_0005s02700	INT	–	–	–	AT3G05420	ACYL-COA BINDING PROTEIN 4 (ACBP4)
Lignin	scaffold_5_8228988	POPTR_0005s11380	INT	+	–	–	AT2G17820	Histidine kinase 1 (ATHK1)
Lignin	scaffold_5_18206992	POPTR_0005s19380	5′-UTR	+	–	–	AT1G10020	Unknown protein
Lignin	scaffold_5_23631116	POPTR_0005s25350	CDS	+	V/F	Non-Syn	AT1G75430	BEL1-LIKE HOMEODOMAIN 11 (BLH11); transcription factor
Lignin	scaffold_5_23631273	POPTR_0005s25350	INT	+	–	–	AT1G75430	BEL1-LIKE HOMEODOMAIN 11 (BLH11); transcription factor
Lignin	scaffold_6_2344494	POPTR_0006s03590	CDS	+	H/R	Non-Syn	AT3G56970	BHLH038; DNA binding/transcription factor
Lignin	scaffold_6_8510688	POPTR_0006s11200	CDS	–	G/G	Syn	AT3G08640	Alphavirus core protein family
Lignin	scaffold_8_2306887	POPTR_0008s04140	CDS	–	N/N	Syn	AT3G07990	Serine carboxypeptidase-like 27 (SCPL27)
Lignin	scaffold_8_2309317	POPTR_0008s04140	NC	–	–	–	AT3G07990	Serine carboxypeptidase-like 27 (SCPL27)
Lignin	scaffold_8_2703267	POPTR_0008s04700	NC	+	–	–	AT3G55070	Unknown protein
Lignin	scaffold_11_8549487	POPTR_0011s07740	INT	–	–	–	AT2G33990	IQ-domain 9 (iqd9); calmodulin binding
Lignin	scaffold_13_14861001	POPTR_0013s14730	3′-UTR	+	–	–	AT5G54490	PINOID-BINDING PROTEIN 1 (PBP1); calcium ion binding/protein binding

ANTH, AP180 N-terminal homology; ATL, Arabidopsis Tóxicos en Levadura; BHLH, basic helix-loop-helix; CDS, coding sequence; ENTH, Epsin N-terminal homology; FtsH, filamentous temperature-sensitive; GCN5, General Control Of Amino-Acid Synthesis, yeast, homolog; Hsf, Heat Stress Transcription Factor; INT, intron; IQ, isoleucine glutamine; LIM: Lin11, Isl-1 and Mec-3; LRR, Leucine-Rich Repeats; na, not applicable; NB-ARC: nucleotide-binding adaptor shared by APAF-1, R proteins, and CED-4; NC, noncoding; non-syn, nonsynonymous; NSF, Soluble N-ethylmaleimide-sensitive factor; SGT1, Suppressor of G2 (Two) allele of skp1; SNAP, Attachment Protein Receptor; syn, synonymous; TUB, tubulin beta chain; UTR, untranslated region; VHS: Vps27, Hrs and STAM.

[1] Pre-selected transcripts with logBF > 0 according to *P. trichocarpa* genome version 2.2 annotations where available, in bold.

[2] Pre-selected SNPs with logBF > 4 given as nucleotide position in *P. trichocarpa* genome version 2.2 scaffolds (www.phytozome.net).

[3] *P. trichocarpa* genome version 2.2 gene model (www.phytozome.net).

Table 3 List of *P. trichocarpa* transcripts validated across different pre-selection analyses: *t*-test with $q < 0.1$ cut-off; Random Forest (RF); logarithm of Bayes factor (logBF)

Gene/transcript	Gene model	Pre-selection analysis	Trait	At ID	Annotation
fgenesh4_pg.C_LG_VI001909	POPTR_0006s28070.1	*t*-test	Lignin, alpha-cellulose, glucose	AT4G30630.1	Unknown protein
		RF	Lignin, alpha-cellulose, glucose		
		logBF	Lignin, alpha-cellulose		
gw1.XVIII.85.1	POPTR_0018s12580.1	*t*-test	Hemicellulose/xylose	AT5G35200.1	ENTH/ANTH/VHS superfamily protein
		logBF	Hemicellulose		
estExt_fgenesh4_pg.C_102780001	POPTR_0001s45360.1	*t*-test	Hemicellulose/xylose	AT1G32900.1	UDP-Glycosyltransferase superfamily protein
		RF	Xylose		
estExt_fgenesh4_pg.C_LG_X0832	POPTR_0010s10140.1	*t*-test	Hemicellulose/xylose	AT2G09990.1	40S ribosomal protein S16 (RPS16A)
		RF	Xylose		
eugene3.00130825	na	*t*-test	Glucose	na	na
		RF	Glucose		
eugene3.00130933	POPTR_0015s02632.1	*t*-test	Alpha-cellulose	AT5G18880.1	Non-LTR retroelement reverse transcriptase related, putative AP endonuclease/reverse transcriptase
		RF	Lignin		
eugene3.00700144	POPTR_0005s04680.1	*t*-test	Alpha-cellulose	na	na
		RF	Glucose		
eugene3.02080032	POPTR_0001s14920.1	*t*-test	Hemicellulose/xylose	AT3G50160.1	Plant protein of unknown function (DUF247)
		RF	Xylose		
eugene3.03180006	POPTR_0125s00200.1	*t*-test	Hemicellulose/xylose	na	na
		RF	Xylose		
eugene3.81200001	POPTR_0005s02810.1	*t*-test	Hemicellulose/xylose	AT5G48930.1	Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (HCT)
		RF	Xylose		
fgenesh4_pg.C_scaffold_7151000001	POPTR_0014s04140.1	*t*-test	Alpha-cellulose	AT4G01070.1	UDP-Glycosyltransferase superfamily protein
		RF	Glucose		
grail3.0001078401	POPTR_0009s07970.1	*t*-test	Hemicellulose/xylose	AT5G58520.1	Protein kinase superfamily protein
		RF	Hemicellulose		
gw1.57.289.1	POPTR_0005s12750.1	*t*-test	Hemicellulose/xylose	AT4G36730.2	G-box binding factor 1 (GBF1); transcription factor
		RF	Xylose		
gw1.I.4726.1	POPTR_0001s37450.1	*t*-test	Lignin	AT5G55600.1	Agenet domain-containing protein/bromo-adjacent homology (BAH) domain-containing protein
		RF	Lignin		
gw1.XVI.2416.1	POPTR_0016s07680.1	*t*-test	Hemicellulose/xylose	AT5G22380.1	Arabidopsis NAC domain containing protein 90 (ANAC090); transcription factor
		RF	Xylose		
estExt_Genewise1_v1.C_LG_III0064	POPTR_0003s15790.1	RF	Glucose	AT3G16560.1	Protein phosphatase 2C-related/PP2C-related
		logBF	Glucose		
eugene3.00150219	POPTR_0015s06200.1	RF	Glucose	na	na
		logBF	Glucose		
eugene3.02220002	POPTR_0015s15100.1	RF	Xylose	AT4G08850.1	Leucine-rich repeat family protein/protein kinase family protein
		logBF	Xylose		
fgenesh4_pg.C_LG_X001551	POPTR_0010s17910.1	RF	Glucose	na	na
		logBF	Glucose, alpha-cellulose		
fgenesh4_pg.C_scaffold_12767000001	POPTR_0009s00200.1	RF	Xylose	na	myb family transcription factor
		logBF	Xylose		
grail3.0005024601	POPTR_0015s10460.1	RF	Hemicellulose	AT5G20090.1	Uncharacterised protein family (UPF0041)
		logBF	Hemicellulose		
grail3.0045020801	POPTR_0004s18330.1	RF	Xylose	na	Putative splicing factor (*Arabidopsis thaliana*)
		logBF	Hemicellulose		
gw1.8759.5.1	POPTR_0017s01420.1	RF	Xylose	AT3G14460.1	LRR and NB-ARC domains-containing disease resistance protein
		logBF	Xylose		

ANTH, AP180 N-terminal homology; AP, Apurinic/apyrimidinic; ENTH, Epsin N-terminal homology; LRR, Leucine-Rich Repeats; LTR, Long terminal repeat; na, not applicable; NB-ARC, nucleotide-binding adaptor shared by APAF-1, R proteins, and CED-4; VHS: Vps27, Hrs and STAM.

percentage of identified SNPs were also identical between hemicellulose content and xylose content traits (*c.* 40%). This was not unexpected as these traits are related: (glucurono-) xylan is the predominant hemicellulose in poplar wood and the two traits are positively correlated (phenotypic correlation $r = 0.80$). Interestingly, however, fewer SNPs (< 14%) were shared between glucose and alpha-cellulose content traits, although most of the cell wall glucose is derived from alpha-cellulose (phenotypic correlation $r = 0.82$). Other sources of glucose in poplar wood include starch, glucomannan, and xyloglucan.

Directed and phenotype-centric networks for six wood attributes

We next used the pre-selected transcripts and SNPs for integrative network analysis as previously described (Chang & McGeachie, 2011) to identify a set of variables with high predictability for the phenotype. Here, the logBF is evaluated for the whole set of modelled networks. This analysis provided complementary information to the individual variable selection procedure described in the previous section above. This integrative Bayesian network analysis has the potential to uncover SNP–SNP, SNP–transcript, transcript–transcript, phenotype–SNP and phenotype–transcript interactions. Within the biological context, SNP–SNP interactions identify loci at higher linkage disequilibrium, SNP–transcript interactions point to expression QTLs, while transcript–transcript interactions identify a functional relationship. It is the intrinsic property of directionality in Bayesian networks (Friedman *et al.*, 2000) that makes this approach useful for phenotype prediction. Thus, this method connects SNPs with phenotypes on a different basis from that of genetic associations. Here, the phenotype represents the root node of the entire network and all nodes within this network are directly or indirectly connected to the phenotype. This allows the prediction of phenotypic values on the basis of values for all other variables accommodated in the network.

Both pre-selected transcripts and SNPs were used to model the SNP–gene networks as trait-centric networks to identify transcripts and SNPs highly associated with the traits under study and to infer directionality of interactions between the variables (Chang & McGeachie, 2011). First, we tested SNPs with very high statistical support for being individually dependent on the phenotype (logBF > 4). For all six wood attributes for which we had conducted the pre-filtering steps, as described in a previous section, network learning steps were performed. The learned networks incorporated many pre-selected SNPs but, surprisingly, no pre-selected transcripts were accommodated in these networks as being directly or indirectly linked to a studied phenotype. Generally, transcripts were pre-selected at a much lower logBF threshold than SNPs, a phenomenon also reported in the original application of this analysis (Chang & McGeachie, 2011). While the differential gene expression described above may actually contribute to phenotypic variation, the small variation in transcript abundance observed for any of these genes does not seem to be an important predictor for the studied phenotypes, contrary to the allelic variation at SNPs. It was also our goal to identify a minimum set of variables from both transcripts and SNPs (if applicable) to achieve the highest possible predictability for the investigated phenotype.

Next, we tested all transcripts and SNPs (using stringent selection criteria; Table 2) in Bayesian network analysis employing the local propagation algorithm to compute the most probable phenotypic value (Chang & McGeachie, 2011). To identify variables with high statistical support for being correlated with the phenotype or with other variables connected to the phenotype, and to test the robustness of all network models, we conducted leave-one-out cross-validation. This approach and its results are summarized in Table 4. In brief, the approach involved removing one individual at a time from the full sample set, computing the networks and each time predicting the three phenotype classes with certain accuracy. For the final networks we used again all 16 genotypes and tested only those variables that were accommodated persistently in the leave-one-out cross-validation approach. As can be seen in Table 4, this approach greatly reduced over-fitting and improved the overall accuracy in predicting the 16 data points (genotypes). The final networks for all six phenotypes are displayed in Figs 2(a–d) and 3(a,b). In all cases, the variables identified in the networks had direct functional links with the studied phenotypes. Except for the glucose-centric network (Fig. 2a), in each phenotype-centric network at least one SNP (in total 13 different SNPs) derived from a candidate gene placed on the SNP array as being potentially related to secondary cell wall biogenesis (Table S8). Examples of such polymorphisms include: scaffold_2_1322346 within POPTR_0002s02340, homologous

Table 4 Cross-validation results for network analysis

	Initial assessment using leave-one-out procedure					Final assessment using full sample set and variables with 100% persistency in the initial assessment		
Trait	No. of genes tested (logBF > 0)	No. of genes accomodated in networks	No. of SNPs tested (logBF ≥ 4)	No. of SNPs accomodated in networks	Predictive accuracy (average) (%)	No. of SNPs tested (logBF ≥ 4)	No. of SNPs accomodated in networks	Predictive accuracy (%)
Alpha-celluose	5	0	6	6	100	6	4	94
Glucose	7	0	7	7	93	6	6	94
Xylose	10	0	7	7	88	6	6	94
Hemicellulose	8	0	16	15	91	8	6	100
Lignin	3	0	20	18	81	6	6	100
Density	1	0	18	18	74	5	5	94

SNP, single nucleotide polymorphism; logBF, logarithm of Bayes factor.

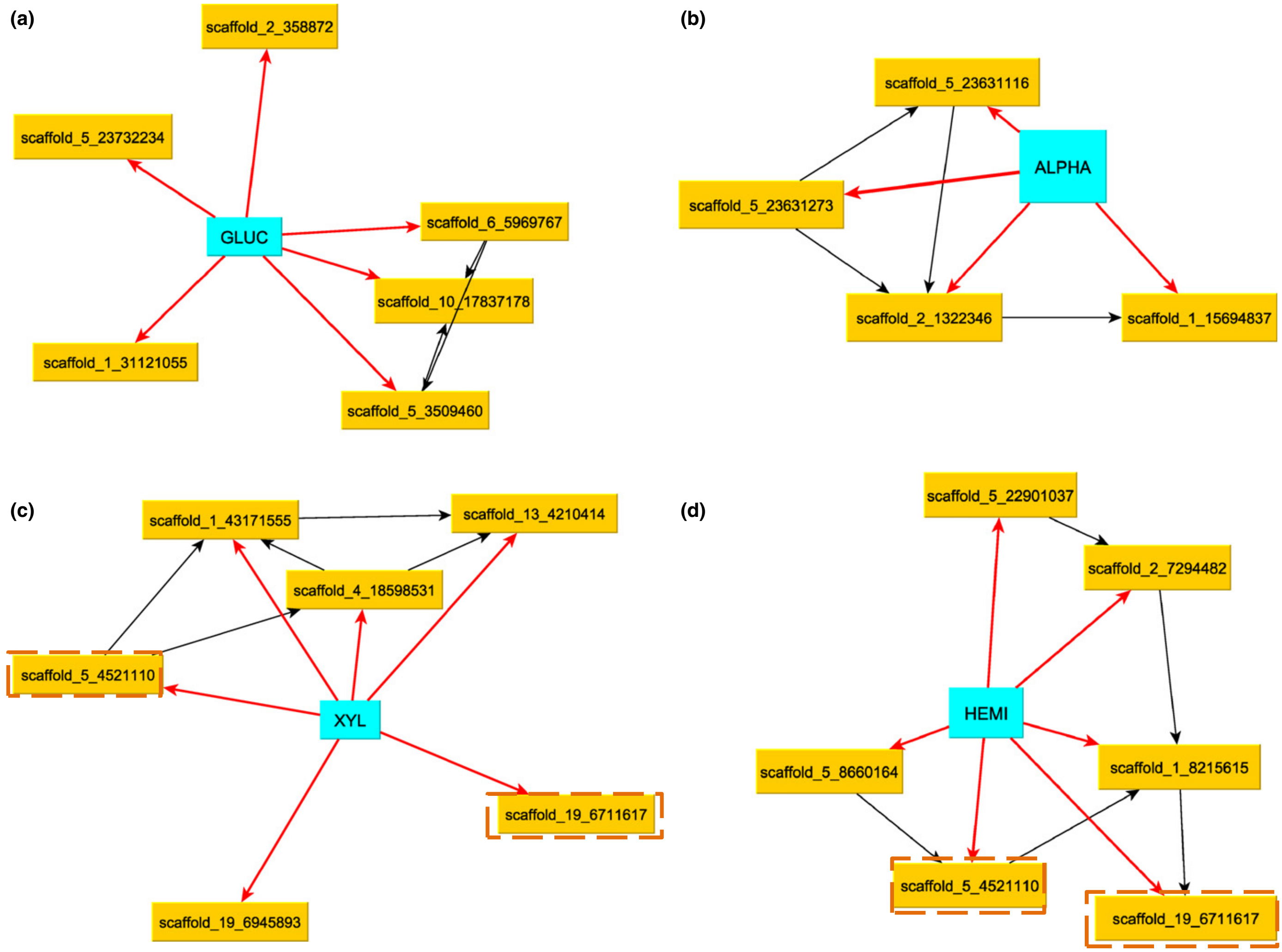

Fig. 2 Representation of the integrative networks for glucose content, alpha-cellulose content, xylose content, and hemicellulose content in *Populus trichocarpa*. Directed networks are shown for (a) the monomeric cell wall sugar glucose (GLUC); (b) the polysaccharide alpha-cellulose (ALPHA); (c) the monomeric cell wall sugar xylose (XYL); and (d) the polysaccharide hemicelluloses (HEMI). A K2 algorithm was employed to learn the best network; the maximum number of parents of each node was restricted to be three; blue filled rectangles indicate the respective phenotype under study; the variables important for phenotype prediction are shown within yellow filled rectangles; arrows point from parent to child node; each connector line represents a directed link between a pair of nodes; in red are direct connections between variables and the studied wood trait; orange dashed boxes indicate variables common between xylose (c) and hemicelluloses (d). Genes represented by each single nucleotide polymorphism (SNP) variable in the networks are given in Table 2.

to the Arabidopsis structural constituent of the cytoskeleton tubulin beta-5 chain (TUB5) (Fig. 2b) (alpha-cellulose); scaffold_1_2936883 within POPTR_0001s03670, a putative cinnamoyl-CoA reductase (CCR) (Fig. 3a) (lignin); and scaffold_1_12114727 within POPTR_0001s15280, homologous to the Arabidopsis GTP-binding RAS 5 (ARA-5) (Fig. 3b) (wood density). In networks such as those for xylose, hemicelluloses and density (Figs 2c,d, 3b), we identified 67% for both xylose and hemicelluloses, and 60% for density, respectively, of the accommodated SNPs from the group of secondary cell wall candidate genes. In two cases the same SNPs were accommodated in networks for different but highly related traits, that is, xylose and hemicelluloses (Fig. 2c,d). These common variables showed similar logBF values for different related traits (see also Table S7).

Gene functions of variables associated with notable differences in secondary cell wall traits

A relatively small number of variables pre-selected at a stringent logBF > 4 were sufficient to accurately predict phenotypic subclasses in the studied population (Table 4). Among these were SNPs within genes encoding microtubule-associated proteins and proteins related to cell wall metabolism (polysaccharide synthesis and cell wall reassembly; see also previous section). One SNP within POPTR_0019s06140, annotated as xyloglucan galactosyltransferase (XGT), GT47 gene family, was highly related to variation in relative xylose content (hemicellulose content). Other genes with expression patterns suggesting involvement in secondary cell wall formation were also identified (Table 2). For example, subtilisins are present in vascular bundles and are involved in

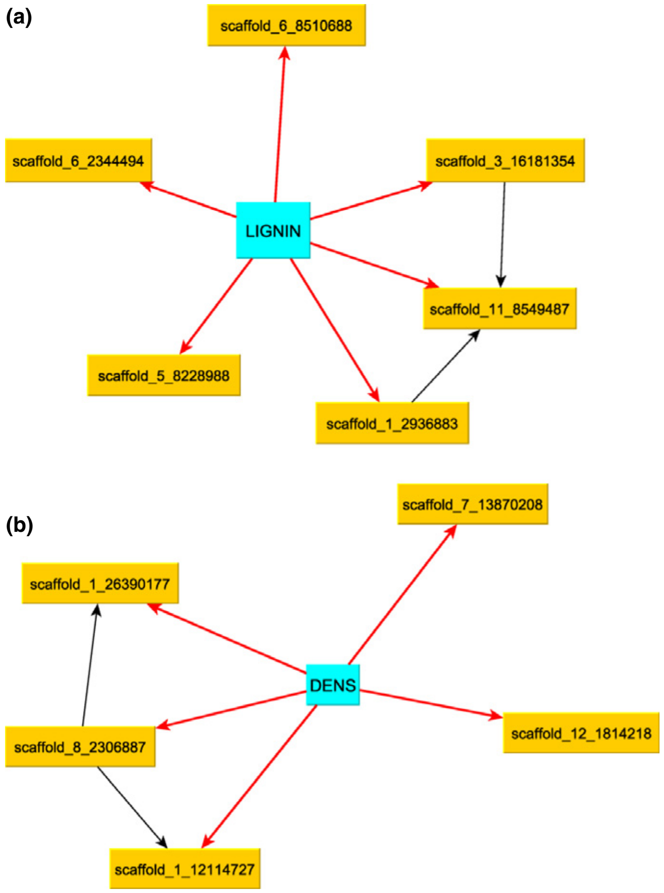

Fig. 3 Representation of the integrative networks for lignin content and average wood density in *Populus trichocarpa*. Directed networks are shown for (a) lignin (LIGNIN) and (b) average wood density (DENS). A K2 algorithm was employed to learn the best network; the maximum number of parents of each node was restricted to be three; blue filled rectangles indicate the respective phenotype under study; the variables important for phenotype prediction are shown within yellow filled rectangles; arrows point from parent to child node; each connector line represents a directed link between a pair of nodes; in red are direct connections between variables and the studied wood trait. Genes represented by each single nucleotide polymorphism (SNP) variable in the networks are given in Table 2.

the modification of cell wall structure and xylem differentiation (Zhao *et al.*, 2005). One SNP located within a gene encoding a subtilase family protein (subtilisin-like serine protease) was identified in our study for relative xylose content (scaffold_1_43171555 within POPTR_0001s43080; Fig. 2c, and Table S8). An SNP within a gene encoding a pectin esterase (for average wood density, scaffold_12_1814218 within POPTR_0012s02170; Fig. 3b) was also identified. Potential roles for pectin esterases in wood density and cell wall thickening have been suggested in *Pinus radiata*, where allelic variation in such genes is associated with variation in this trait (Li *et al.*, 2012).

Finally, pre-selected variables for variation in relative total lignin content included numerous SNPs within genes encoding proteins required for monolignol biosynthesis. These included poplar genes with homology to Arabidopsis genes encoding prephenate dehydratase and arogenate dehydratase (*PD1* and *ADT1*; shikimate pathway), cinnamoyl CoA reductase (cinnamoyl-CoA reductase 1 (*CCR1*)) and caffeic acid O-methyltransferase (caffeic acid O-methyltransferase 2 (*COMT2*); phenylpropanoid pathway), and poplar genes with homology to Arabidopsis genes encoding proteins potentially involved in monolignol polymerization (*IRREGULAR XYLEM12* (*IRX12*) orthologues *LACCASE17* (*LAC17*) and *LAC3*) as well as potential transcriptional regulators for secondary growth (*WRKY35*, *MYB63* and *WRKY70* (Geraldes *et al.*, 2013); Table S8). Interestingly, with the exception of *CCR1*, none of these genes were required in the SNP–SNP network to accurately predict strong variation in lignin content (Fig. 3a).

Eight polymorphisms identified in our pre-selection panel (Table S7) were also identified in a genome-wide association study for wood chemistry and ultrastructural traits in a *P. trichocarpa* association population of 334 individuals that included the accessions with extreme phenotypes from the current study (Porth *et al.*, 2013a). The SNP scaffold_2_18793288 within POPTR_0002s22090 which encodes an auxin-responsive family protein was important for explaining variation in both cellulose microfibril angle in the association study (Porth *et al.*, 2013a) and density in this study (Table S8). In addition, the SNP scaffold_14_3194356 within POPTR_0014s04020, annotated as homologous to the *ARABIDOPSIS THAUMATIN-LIKE PROTEIN3* (*ATLP3*) gene, explained variation in related traits (S-lignin in the association study and total lignin in this study) and SNPs within POPTR_0015s05540, homologous to *Arabidopsis SUCROSE SYNTHASE6* (*SUS6*), explained variation for holocellulose in the association study and for glucose content in this study. Importantly, the nonsynonymous SNP scaffold_9_6687336 within POPTR_0009s07050, which encodes a homologue of Arabidopsis MITOGEN-ACTIVATED PROTEIN KINASE 3 (MPK3), was pre-selected for association with glucose content in this study and was associated with the same trait in the association study of Porth *et al.* (2013a).

Discussion

Genetic complexity of wood attributes and predictive modelling of phenotype variation

Many attributes related to wood and secondary cell wall chemistry and ultrastructure (e.g. microfibril angle orientation and wood density) are of polygenic genetic architecture in trees (Li *et al.*, 2011, 2012). In this study, we focused on identifying DNA structural variants (SNPs) and transcript levels that may underlie the phenotypic variation in selected wood chemistry traits and wood density that we previously observed (Porth *et al.*, 2013b), using a Bayesian network approach. We generated integrative networks of these variables that are highly predictive of the phenotype subclasses for each of the six industrially important wood traits studied.

The anticipated complexity of the traits we investigated is reflected in highly complex differences in transcript abundance between individuals with contrasting phenotypes. This is evident in the large number of transcripts with relatively small expression changes that characterized the differentially expressed transcriptomes of poplar individuals with substantial differences in wood attributes. The functional annotation of differentially expressed genes showed that such genes were associated with many distinct molecular functions (results not shown). The complexity of polygenic trait architecture highlights the challenges for the genetic improvement of economically important traits in tree breeding strategies. First, this makes it difficult to identify and capture the entire heritable portion of the observed phenotypic variance in such traits using quantitative genetics approaches such as QTL mapping (in pedigrees) or association mapping (in unstructured wild populations). Secondly, establishing individual trees combining all valuable allelic variants through directed breeding is cumbersome, when large numbers of allelic combinations for multiple traits have to be captured. The genome-wide approach of simultaneously fitting unbiased genetic markers to phenotype expression was recently shown to capture the majority of the genetic variation underlying wood traits and to provide relatively high predictive accuracy for such polygenic traits in tree breeding populations (Resende *et al.*, 2012). However, to infer the *causal* relationships among genes that control the observed gene expression patterns, other approaches have been developed. For example, more recently an integrative network analysis that combines SNP variation and gene expression data was proposed by Chang & McGeachie (2011). In the current study, we used this Bayesian network learning approach which integrates both the genotypic and gene expression data to predict the phenotype under study.

Networks of allelic variants achieve the best predictive accuracy for considerable natural variation in poplar wood traits

A novel aspect of our study is the utilization of natural poplar accessions at the extremes of phenotypic trait values found across a large unstructured population (classified into discrete 'high', 'medium' and 'low' classes; Porth *et al.* (2013b), to rigorously screen tens of thousands of molecular variables for their functional association with these attributes. Our goal was to identify a *minimal* set of variables (representing variation in gene expression and allelic variants) that has the potential of accurately predicting these phenotypic changes and to obtain insight into the hierarchy of the genetic architecture of the studied traits. While logBF preselection identified a number of transcripts that were significantly dependent on a given phenotype (Tables 1–3), in the final networks no single transcript was accommodated. Instead, the networks were strongly dependent on SNP variants. These results suggest that the observed phenotype-dependent transcript variation is not essential in the final network topology to achieve 100% predictability of the phenotype. To validate this finding, we tested predictabilities using the Random Forest algorithm optimized for microarray data (Diaz-Uriarte & de Andres, 2006). Although the selected sets of transcripts reflected well results from the *t*-tests (Fig. S1; Table 3), they provided overall only low predictability for each phenotype, with error rates of 0.415 on average. This general result might be attributable to the existence of numerous transcripts with relatively small effects on natural trait variation, as also shown for other wood traits (Li *et al.*, 2011, 2012), or major effect changes but in different genes may account for the variation. Such inconsistent expression changes would lead to poor predictive ability. The method introduced by Chang & McGeachie (2011) has the potential to identify *cis*-eQTLs, which represent expression QTLs caused by polymorphisms within the gene queried for expression variation (cf. Porth *et al.*, 2011, 2012). However, the makeup of the 34k SNP array (based on broad-based candidate genes for secondary growth; Geraldes *et al.*, 2013) was such that 5% of its genes were not represented on the Nimblegen array, while 90% of the poplar genes on the gene expression array were not considered for the SNP chip. Still, we had common representation for 3366 gene models to test for *cis*-eQTLs. However, in order to identify transcripts, expression changes would have to be fixed within genotypes at the tails of the population distribution. Moreover, SNPs within coding regions can affect gene function without any requirement for expression level changes, while expression level changes probably result from SNPs in promoter regions and not in the coding sequence.

A striking result of our study of six different wood attributes was that, in all cases, SNP–SNP combinations assembled the optimal signatures for 100% or close to 100% predictive accuracies (Table 4). In two cases (hemicelluloses and lignin), we showed that networks with fewer variables but with very stringent selection criteria regarding the individual association of variables with phenotypic variation (logBF > 4) were sufficient to accurately predict the different phenotype subclasses (Figs 2d, 3a). In particular, in the cases of relative hemicellulose content, relative total lignin content, and average wood density, we see how variable reduction improves the predictive accuracy. Thus, from these analyses it appears that cross-validation using the leave-one-out approach is indispensible to avoid model over-fitting and achieve the best predictive accuracy for the final network analysis.

Interestingly, the glucose-centric content network (Fig. 2a) revealed an important direct dependence on SNP variants in genes for biomass accumulation involving POPTR_0005s25490, a peptide transporter expressed in the vascular tissue, POPTR_0010s20100, encoding a putative transcription regulator and a homologue of LIM (Lin11, Isl-1 and Mec-3) domain protein 1b from *Nicotiana tabacum* (NtLIM1), a Pal-box binding protein found in phenylpropanoid genes (Kawaoka *et al.*, 2000), and finally POPTR_0001s32810, a gene homologous to Arabidopsis *HEAT SHOCK TRANSCRIPTION FACTOR A5* (*HSFA5*), a nonconventional heat-shock transcription factor gene that functions as a repressor in the HSFA4/HSFA5 complex and as such may control cell-type-specific apoptosis (Baniwal *et al.*, 2007). All three variants were directly dependent on relative glucose content, implying tight regulation. This highlights the complex interaction between cell wall components and developmental or environmental stimuli in shaping secondary cell wall phenotypes, and is illustrated by these three SNPs. The identified gene hierarchy in this network (Fig. 2a) may also provide indications of why transgenic plants with misregulated genes in a cell wall metabolic pathway often have dramatic perturbations in related metabolic pathways and in overall plant physiology (Coleman *et al.*, 2008; Kitin *et al.*, 2010; Voelker *et al.*, 2011; Vanholme *et al.*, 2012).

In addition, we uncovered important effects for allelic variants within putative poplar *TUB* and *CCR* genes (scaffold_2_1322346 and scaffold_1_2936883, respectively) that were important predictors within the relative alpha-cellulose content and relative lignin content trait networks, respectively (Figs 2b, 3a). CCR is an important enzyme in the metabolic pathway leading to monolignol biosynthesis (Vanholme *et al.*, 2012), and the identified CCR is highly similar (> 92% sequence identity) to the poplar CCR that resulted in reduced lignin, reduced hemicellulose, and increased cellulose upon down-regulation (Leple *et al.*, 2007). Certain poplar beta-tubulins (TUBs) are components of microtubules that influence cellulose deposition in secondary walls (Oakley *et al.*, 2007) and also might affect structural cellulosic wood traits such as microfibril angle (Spokevicius *et al.*, 2007). Thus, we confirmed that certain biosynthetic genes and genes encoding proteins for structural properties can affect substantial natural variation in wood traits. We also found evidence for phytohormone signalling of plant growth and development as components in some of the trait networks. For example, within the density network, an SNP in POPTR_0001s15280 (encoding Rab GTPase *ARA5*) is directly dependent on an SNP within POPTR_0008s04140, annotated as a homologue of the Arabidopsis serine carboxypeptidase BRASSINOSTEROID INSENSITIVE1 SUPRESSOR1 (BRS1) which influences brassinosteroid signalling in Arabidopsis (Li *et al.*, 2001). The poplar homologue of *ARA5* is in turn highly co-regulated with secondary cell wall *CELLULOSE SYNTHASE A* (*CESA*) genes (Persson *et al.*, 2005).

Finally, a poplar NAM, ATAF1,2, CUC2 (NAC) domain transcription factor homologue of Arabidopsis NAC protein 036 (ANAC036) (POPTR_0004s18860) was found in the xylose-centric network. While the specific function of *ANAC036* is unknown, NAC transcription factors such as SECONDARY WALL-ASSOCIATED NAC DOMAIN PROTEIN1 (SND1), VASCULAR-RELATED NAC-DOMAIN7 (VND6) and *VND7* play key roles as regulators of secondary wall formation in vessels and fibres (Wang & Dixon, 2012). However, to date no specific regulator of xylan biosynthesis is known and our findings suggest that POPTR_0004s18860 could represent such a candidate in poplar that can be further explored, at the very least to confirm pleiotropic effects for this transcription factor that can impact the amount of xylan in the wood and be of potential use for manipulating xylan levels in woody biomass (Petersen *et al.*, 2012).

Conclusions

Many complex traits, such as secondary cell wall composition, are the products of multiple highly (positively and/or negatively) correlated and interacting traits. QTL mapping and association mapping studies have provided insight into the genetic architecture of quantitative traits by bridging the gap between phenotypic and genotypic variances. Often, the reported significant associations attributable to each SNP show a small amount (0.1–5%) of the total variance and the effects of QTL are highly dependent on the genetic background in the studied population (Mackay *et al.*, 2009). While large-effect QTLs are found in populations based on crosses of highly divergent parents, small-effect QTLs are reported in populations with large sample sizes. Also, interaction networks are not persistent and may change as breeding advances, especially in forest trees as a consequence of their long-term exposure to large environmental contingencies that can drastically change the realized QTL effects. This manifests in reduced accuracy of genomic selection (Resende *et al.*, 2012). Therefore, the effect of each SNP is not a straightforward criterion according to observed association with the trait under study, especially in a significant context-dependent scenario (Mackay *et al.*, 2009). Genetic network analysis is expected to unravel and detect hierarchies in gene action leading to the discovery of the most important genes affecting the trait in question. The network approach may facilitate the selection of highly important genes and gene variants involved in controlling a specific trait in the complexity of these interacting traits (Porth *et al.*, 2013b). Focus on such gene variants in cell wall traits could simplify breeding activities, as much of the attention would be focused on genotype–phenotype associations across all studied traits that are most likely to result in appreciable improvement by supporting the maximization of response to selection.

Acknowledgements

We acknowledge C. Hefer, UBC, Canada, for updating Nimblegen array annotation to v2 poplar gene models, and A. Haegert at Jack Bell Research Centre, Prostate Centre, Vancouver, BC, Canada for microarray hybridizations and basic bioinformatics analysis. We further thank Drs H. H. Chang and M. McGeachie, Harvard Medical School, MA, USA for MATLAB scripts. This work was supported by the Genome British Columbia Applied Genomics Innovation programme (Project 103BIO) and the Genome Canada Large-Scale Applied Research Project programme (Project 168BIO).

References

Baniwal SK, Chan KY, Scharf K-D, Nover L. 2007. Role of heat stress transcription factor HsfA5 as specific repressor of HsfA4. *Journal of Biological Chemistry* 282: 3605–3613.

Bassel GW, Gaudinier A, Brady SM, Hennig L, Rhee SY, De Smet I. 2012. Systems analysis of plant functional, transcriptional, physical interaction, and metabolic networks. *Plant Cell* 24: 3859–3875.

Beaulieu J, Doerksen T, Boyle B, Clement S, Deslauriers M, Beauseigle S, Blais S, Poulin P-L, Lenz P, Caron S *et al.* 2011. Association genetics of wood physical traits in the conifer White Spruce and relationships with gene expression. *Genetics* 188: 197–214.

Breiman L. 2001. Random forests. *Machine Learning* 45: 5–32.

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P. 2005. Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology* 28: 171–182.

Chang H-H, McGeachie M. 2011. Phenotype prediction by integrative network analysis of SNP and gene expression microarrays. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 6849–6852.

Coleman HD, Samuels AL, Guy RD, Mansfield SD. 2008. Perturbed lignification impacts tree growth in hybrid Poplar-A function of sink strength, vascular integrity, and photosynthetic assimilation. *Plant Physiology* **148**: 1229–1237.

Demura T, Ye Z-H. 2010. Regulation of plant biomass production. *Current Opinion in Plant Biology* **13**: 299–304.

Diaz-Uriarte R, de Andres SA. 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**: 3.

Dinus RJ. 2000. Genetic modification of short rotation poplar biomass feedstock for efficient conversion to ethanol: bioenergy feedstock development program, environmental sciences division. *Oak Ridge National Laboratory* **6**: 7.

Efron B, Tibshirani R. 1997. Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* **92**: 548–560.

Ferrazzi F, Sebastiani P, Ramoni MF, Bellazzi R. 2007. Bayesian approaches to reverse engineer cellular systems: a simulation study on nonlinear Gaussian networks. *BMC Bioinformatics* **8**: S2.

Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7**: 601–620.

Geraldes A, Difazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N *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. *Molecular Ecology Resources* **13**: 306–323.

Geraldes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann M, Birol I *et al.* 2011. SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Molecular Ecology Resources* **11**: 81–92.

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

Harfouche A, Meilan R, Kirst M, Morgante M, Boerjan W, Sabatti M, Mugnozza GS. 2012. Accelerating the domestication of forest trees in a changing world. *Trends in Plant Science* **17**: 64–72.

Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlen M, Teeri TT, Lundeberg J *et al.* 2001. A transcriptional roadmap to wood formation. *Proceedings of the National Academy of Sciences, USA* **98**: 14732–14737.

Hon DNS. 1981. Yellowing of modern papers. In: Williams JC, ed. *Preservation of paper and textiles of historic and artistic value II.* Washington, DC, USA: American Chemical Society, 119–141.

Ingvarsson PK, Street NR. 2011. Association genetics of complex traits in plants. *New Phytologist* **189**: 909–922.

Kawaoka A, Kaothien P, Yoshida K, Endo S, Yamada K, Ebinuma H. 2000. Functional analysis of tobacco LIM protein Ntlim1 involved in lignin biosynthesis. *Plant Journal* **22**: 289–301.

Kirst M, Myburg AA, De Leon JPG, Kirst ME, Scott J, Sederoff R. 2004. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. *Plant Physiology* **135**: 2368–2378.

Kitin P, Voelker SL, Meinzer FC, Beeckman H, Strauss SH, Lachenbruch B. 2010. Tyloses and phenolic deposits in xylem vessels impede water transport in low-lignin transgenic poplars: a study by cryo-fluorescence microscopy. *Plant Physiology* **154**: 887–898.

Leple J-C, Dauwe R, Morreel K, Storme V, Lapierre C, Pollet B, Naumann A, Kang K-Y, Kim H, Ruel K *et al.* 2007. Downregulation of cinnamoyl-coenzyme a reductase in poplar: multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell* **19**: 3669–3691.

Li J, Lease KA, Tax FE, Walker JC. 2001. BRS1, a serine carboxypeptidase, regulates BRI1 signaling in *Arabidopsis thaliana. Proceedings of the National Academy of Sciences, USA* **98**: 5916–5921.

Li X, Wu HX, Southerton SG. 2011. Transcriptome profiling of *Pinus radiata* juvenile wood with contrasting stiffness identifies putative candidate genes involved in microfibril orientation and cell wall mechanics. *BMC Genomics* **12**: 480.

Li X, Wu HX, Southerton SG. 2012. Identification of putative candidate genes for juvenile wood density in *Pinus radiata. Tree Physiology* **32**: 1046–1057.

Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. 2004. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics* **5**: 32.

Ma S, Gong Q, Bohnert HJ. 2007. An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research* **17**: 1614–1625.

Mackay TFC, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**: 565–577.

Martin A, Quinn K. 2007. *MCMCpack: Markov chain Monte Carlo (MCMC) Package.* R package version 0.8-1, URL: http://mcmcpack.wustl.edu.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. *Nature Reviews Genetics* **12**: 111–122.

Nieminen K, Robischon M, Immanen J, Helariutta Y. 2012. Towards optimizing wood development in bioenergy trees. *New Phytologist* **194**: 46–53.

Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CRD, Benedict C, Dervinis C, Yu Q, Sykes R, Davis M *et al.* 2009. Quantitative genetic analysis of biomass and wood chemistry of *Populus* under different nitrogen levels. *New Phytologist* **182**: 878–890.

Oakley RV, Wang Y-S, Ramakrishna W, Harding SA, Tsai C-J. 2007. Differential expansion and expression of alpha- and beta-tubulin gene families in *Populus. Plant Physiology* **145**: 961–973.

Persson S, Wei HR, Milne J, Page GP, Somerville CR. 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences, USA* **102**: 8633–8638.

Petersen PD, Lau J, Ebert B, Yang F, Verhertbruggen Y, Kim JS, Varanasi P, Suttangkakul A, Auer M, Loque D *et al.* 2012. Engineering of plants with improved properties as biofuels feedstocks by vessel-specific complementation of xylan biosynthesis mutants. *Biotechnology for Biofuels* **5**: 84.

Plomion C, Leprovost G, Stokes A. 2001. Wood formation in trees. *Plant Physiology* **127**: 1513–1523.

Porth I, Hamberger B, White R, Ritland K. 2011. Defense mechanisms against herbivory in Picea: sequence evolution and expression regulation of gene family members in the phenylpropanoid pathway. *BMC Genomics* **12**: 608.

Porth I, Klápště J, Skyba O, Hannemann J, McKown AD, Guy RD, DiFazio SP, Muchero W, Ranjan P, Tuskan GA *et al.* 2013a. Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate SNPs. *New Phytologist.* doi: 10.1111/nph.12422.

Porth I, Klápště J, Skyba O, Lai BS, Geraldes A, Muchero W, Tuskan GA, Douglas CJ, El-Kassaby YA, Mansfield SD. 2013b. *Populus trichocarpa* cell wall chemistry and ultrastructure trait variation, genetic control and genetic correlations. *New Phytologist* **197**: 777–790.

Porth I, White R, Jaquish B, Alfaro R, Ritland C, Ritland K. 2012. Genetical genomics identifies the genetic architecture for growth and weevil resistance in spruce. *PLoS ONE* **7**: e44397.

Pot D, Chantre G, Rozenberg P, Rodrigues JC, Jones GL, Pereira H, Hannrup B, Cahalan C, Plomion C. 2002. Genetic control of pulp and timber properties in maritime pine (*Pinus pinaster* Ait.). *Annals of Forest Science* **59**: 563–575.

Resende MDV, Resende MFR, . Jr, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM *et al.* 2012. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytologist* **194**: 116–128.

Schrader J, Nilsson J, Mellerowicz E, Berglund A, Nilsson P, Hertzberg M, Sandberg G. 2004. A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *Plant Cell* **16**: 2278–2292.

Spokevicius AV, Southerton SG, MacMillan CP, Qiu D, Gan S, Tibbits JFG, Moran GF, Bossinger G. 2007. Beta-tubulin affects cellulose microfibril orientation in plant secondary fibre cell walls. *Plant Journal* **51**: 717–726.

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

Thumma BR, Matheson BA, Zhang D, Meeske C, Meder R, Downes GM, Southerton SG. 2009. Identification of a *cis*-acting regulatory polymorphism in a Eucalypt *COBRA*-like gene affecting cellulose content. *Genetics* **183**: 1153–1164.

Thumma BR, Southerton SG, Bell JC, Owen JV, Henery ML, Moran GF. 2010. Quantitative trait locus (QTL) analysis of wood quality traits in *Eucalyptus nitens. Tree Genetics & Genomes* **6**: 305–317.

Tsai C-J, Ranjan P, DiFazio SP, Tuskan GA, Johnson V. 2011. Poplar genome microarrays. In: Joshi CP, DiFazio SP, Kole C, eds. *Genetics, genomics and breeding of poplar.* Enfield, NH, USA: Science Publishers, 112–117.

Ursache R, Nieminen K, Helariutta Y. 2012. Genetic and hormonal regulation of cambial development. *Physiologia Plantarum* 147: 36–45.

Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. 2010. Lignin biosynthesis and structure. *Plant Physiology* 153: 895–905.

Vanholme R, Storme V, Vanholme B, Sundin L, Christensen JH, Goeminne G, Halpin C, Rohde A, Morreel K, Boerjan W. 2012. A systems biology view of responses to lignin biosynthesis perturbations in Arabidopsis. *The Plant Cell* 24:3506–3529.

Voelker SL, Lachenbruch B, Meinzer FC, Kitin P, Strauss SH. 2011. Transgenic poplars with reduced lignin show impaired xylem conductivity, growth efficiency and survival. *Plant, Cell & Environment* 34: 655–668.

Wang H-Z, Dixon RA. 2012. On-off switches for secondary cell wall biosynthesis. *Molecular Plant* 5: 297–303.

Wegrzyn JL, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R, Davis MF, Tsai C-J, Neale DB. 2010. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. *New Phytologist* 188: 515–532.

Wimmer V, Albrecht T, Auinger HJ, Schön CC. 2012. Synbreed: a framework for the analysis of genomic prediction data using R. *Bioinformatics* 28: 2086–2087.

Xie C-Y, Ying CC, Yanchuk AD, Holowachuk DL. 2009. Ecotypic mode of regional differentiation caused by restricted gene migration: a case in black cottonwood (*Populus trichocarpa*) along the Pacific Northwest coast. *Canadian Journal of Forest Research* 39: 519–526.

Zhao CS, Craig JC, Petzold HE, Dickerman AW, Beers EP. 2005. The xylem and phloem transcriptomes from secondary tissues of the Arabidopsis root-hypocotyl. *Plant Physiology* 138: 803–818.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Representation of variable (transcript) overlaps with comparison of *t*-test results ($P < 0.05$, $FC > 1.2$), Random Forest (RF), and logarithm of Bayes Factor (logBF) variable selection procedures.

Table S1 Range of phenotypic values in the studied *P. trichocarpa* individuals

Table S2 Experimental setup for *P. trichocarpa* sample hybridizations using the 49K Nimblegen poplar gene expression microarray

Table S3 Re-annotation of array features on the 49K Nimblegen poplar gene expression microarray according to version 2.2 of the *Populus trichocarpa* genome assembly and annotation

Table S4 Input file for integrative network analysis containing information about phenotype subclass categorization, normalized gene expression levels and SNP genotype data for *P. trichocarpa* individuals

Table S5(1–10) Comprehensive results of genes showing differences in developing xylem gene expression ($P < 0.05$; $FC > 1.2$) between *Populus trichocarpa* trees in contrasting trait bins (high vs low)

Table S6 Summary results of *P. trichocarpa* transcripts (genes) identified via *t*-tests using false discovery rate (FDR: $q < 0.1$), Random Forest (RF) clustering approach, and logBF regression approach

Table S7 Significant SNPs with $\log BF > 2$ selection criteria for each phenotype in *P. trichocarpa*

Table S8 Significant SNP variables ($\log BF > 2$) identified by preselection as dependent on phenotype, representing a summary of SNPs within *P. trichocarpa* candidate genes with suggested involvement in secondary cell wall formation (Geraldes *et al.*, 2013)

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