Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms

Ilga Porth[1]*, Jaroslav Klapšte[2,3]*, Oleksandr Skyba[1], Jan Hannemann[4], Athena D. McKown[2], Robert D. Guy[2], Stephen P. DiFazio[5], Wellington Muchero[6], Priya Ranjan[6], Gerald A. Tuskan[6], Michael C. Friedmann[7], Juergen Ehlting[4], Quentin C. B. Cronk[7], Yousry A. El-Kassaby[2], Carl J. Douglas[7] and Shawn D. Mansfield[1]

[1]Department of Wood Science, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4; [2]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4; [3]Department of Dendrology and Forest Tree Breeding, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences, Prague, 165 21, Czech Republic; [4]Department of Biology and Centre for Forest Biology, University of Victoria, Victoria, BC, Canada, V8W 3N5; [5]Department of Biology, West Virginia University, Morgantown, WV 26506-6057, USA; [6]BioSciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA; [7]Department of Botany, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4

Authors for correspondence:
Shawn D. Mansfield
Tel: +1 604 822 0196
Email: shawn.mansfield@ubc.ca

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

Received: *30 March 2013*
Accepted: *18 June 2013*

New Phytologist (2013) **200:** 710–726
doi: 10.1111/nph.12422

Key words: association genetics, cellulose, lignin, *Populus*, single nucleotide polymorphism (SNP), wood traits, wood ultrastructure.

Summary

- Establishing links between phenotypes and molecular variants is of central importance to accelerate genetic improvement of economically important plant species. Our work represents the first genome-wide association study to the inherently complex and currently poorly understood genetic architecture of industrially relevant wood traits.
- Here, we employed an Illumina Infinium 34K single nucleotide polymorphism (SNP) genotyping array that generated 29 233 high-quality SNPs in *c.* 3500 broad-based candidate genes within a population of 334 unrelated *Populus trichocarpa* individuals to establish genome-wide associations.
- The analysis revealed 141 significant SNPs ($\alpha \leq 0.05$) associated with 16 wood chemistry/ultrastructure traits, individually explaining 3–7% of the phenotypic variance. A large set of associations (41% of all hits) occurred in candidate genes preselected for their suggested *a priori* involvement with secondary growth. For example, an allelic variant in the *FRA8* ortholog explained 21% of the total genetic variance in fiber length, when the trait's heritability estimate was considered. The remaining associations identified SNPs in genes not previously implicated in wood or secondary wall formation.
- Our findings provide unique insights into wood trait architecture and support efforts for population improvement based on desirable allelic variants.

Introduction

Plant secondary cell walls that form the bulk of biomass from wood or other lignocellulosic feedstocks are naturally recalcitrant to deconstruction and sugar release (saccharification). Such recalcitrance is attributed to the heterogeneity of the cell wall components, protection of cellulose microfibrils by lignin and hemicelluloses, inaccessibility of enzymes to the cellulose moieties, and cellulose crystallinity (Mansfield *et al.*, 1999; Mosier *et al.*, 2005). Hence, tailoring the lignocellulosic substrate to improve cell wall deconstruction and increase saccharification efficiency for the conversion into liquid biofuels remains an active field of research (Yuan *et al.*, 2008; Carroll & Somerville, 2009; Shen *et al.*, 2012).

Generally, wood from the genus *Populus* (poplars, cottonwoods and aspens) exhibits favorable relative proportions and structures of all three major cell wall components – cellulose, lignin and hemicellulose (Chang & Holtzapple, 2000; Dinus, 2000; Sannigrahi *et al.*, 2010; Studer *et al.*, 2011), with glucan content ranging from 39 to 49% of dry weight (DW), hemicelluloses content from 16 to 23% of DW, and lignin content from 21 to 29% of DW. Although xylan and lignin can serve as precursors for various biomaterials (Hatakeyama & Hatakeyama, 2010; Sannigrahi *et al.*, 2010), for increased ethanol yield, feedstocks with lower lignin and, to some degree, lower xylan content, and an above-average syringyl lignin composition (syringyl : guaiacyl) remain research targets (Studer *et al.*, 2011). Native *Populus* cellulose is *c.* 45–50% crystalline and this intrinsic ultrastructural property is an important parameter in the overall efficiency of the enzymatic

*These authors contributed equally to this work.

saccharification reactions (Mansfield *et al.*, 1999; Chang & Holtzapple, 2000). There is substantial natural variation for all of these properties in *Populus* (R. Kumar *et al.*, 2009; Porth *et al.*, 2013), and this may be exploited through breeding and biotechnological approaches to improve feedstock quality.

Previous studies have aimed at identifying allelic variants in candidate genes that could affect wood phenotypes through quantitative trait locus (QTL) analysis (Novaes *et al.*, 2009; Thumma *et al.*, 2010), or have employed association mapping to 'unstructured' populations to identify significant trait associations with small numbers of candidate genes (Thumma *et al.*, 2009; Wegrzyn *et al.*, 2010; Guerra *et al.*, 2013). Unlike QTL studies, association genetics identifies genetic polymorphisms with higher resolution (as a result of rapid decline of linkage disequilibrium; Rafalski, 2010), which can be exploited for preselection of individuals for tree improvement. Generally, individual single nucleotide polymorphisms (SNPs) typically explain only a small portion of the phenotypic variance ($\leq 5\%$). This has been attributed to the 'complexity' of the traits (Wegrzyn *et al.*, 2010), or to the application of single locus approaches to the multidimensionality of traits (Kremer, 2011). Genome-wide association (GWA) mapping, which has not yet been applied to trees, is also feasible where rich genomic resources enable high-density SNP genotyping (Brachi *et al.*, 2011).

In addition to their economic importance for both traditional and bioenergy applications (Carroll & Somerville, 2009; Sannigrahi *et al.*, 2010; Stanton *et al.*, 2010), trees of the genus *Populus* play important ecological roles in natural ecosystems (Eckenwalder, 1996; Cronk, 2005). Extensive genetic and genomic resources exist for black cottonwood (*Populus trichocarpa*), including a whole-genome reference (Nisqually-1; Tuskan *et al.*, 2006), comprehensive expressed sequence tag (EST) resources, full-length cDNA and SNP databases (Tuskan *et al.*, 2006; Ralph *et al.*, 2008; Geraldes *et al.*, 2011; Slavov *et al.*, 2012), and range-wide collections of wild accessions (Xie *et al.*, 2009; Slavov *et al.*, 2012). Several previous population genomics studies demonstrated that *Populus* has excellent characteristics for association analysis. *Populus* spp. have highly variable linkage disequilibrium (LD) across the genome, relatively weak neutral population structure, extensive geographically structured phenotypic and adaptive genetic variation, and moderate neutral genetic variation (Ingvarsson, 2005; Olson *et al.*, 2010; Keller *et al.*, 2012; Slavov *et al.*, 2012).

Although variations in traits such as wood composition (Ingvarsson *et al.*, 2008; Sannigrahi *et al.*, 2010; Wegrzyn *et al.*, 2010) are well established in wild *Populus* populations, there are few association mapping studies for these traits, and those that do exist are based on genetic variants at a small number of genes (e.g. Ingvarsson *et al.*, 2008; Wegrzyn *et al.*, 2010). We have found that unrelated individuals in a range-wide collection of *P. trichocarpa* grown in a common garden display high degrees of heritable phenotypic variability in wood chemistry and ultrastructure traits (Porth *et al.*, 2013), despite relatively low population genetic structure in this species (Wegrzyn *et al.*, 2010; Slavov *et al.*, 2012; Geraldes *et al.*, 2013). Thus, in this study, we used a 34K SNP genotyping array (Geraldes *et al.*, 2013) and association mapping to identify *P. trichocarpa* allelic variants that underlie variation in key wood chemistry and ultrastructure traits important for bioenergy and traditional uses. Using this array, we identified 141 significant SNPs associated with cell wall traits.

Materials and Methods

Association mapping population and generation of wood phenotypic data

The *Populus trichocarpa* Torr. & A. Gray (black cottonwood) association mapping population we employed was described previously (Xie *et al.*, 2009; Geraldes *et al.*, 2011, 2013; Porth *et al.*, 2013). Briefly, we used increment cores taken from 384 9-yr-old trees from the collection grown in a common garden at Surrey, BC, Canada (Supporting Information, Table S1, Fig. S1).

The phenotyping pipeline used for analysis of wood chemical composition and wood physical and ultrastructural properties is described in Porth *et al.* (2013). In brief, phenotypic variation for 17 wood traits was assessed involving the relative amounts of monomeric cell wall sugars (arabinose, rhamnose, galactose, glucose, xylose, and mannose), polysaccharides (alpha cellulose, holocellulose), and lignin (acid-soluble and -insoluble fractions, as well as their composite total lignin) in dry wood, and relative lignin composition (% syringyl monomers), as well as the solid wood properties, % cell wall crystallinity, average wood density, microfibril angles at first and most recent growth ring, and fiber length. We evaluated the distributions of variation in all traits in the association population and found that all provided normal distribution, some after appropriate data transformation.

SNP genotyping

A detailed description of the selection of the 3543 candidate genes on the 34K Illumina Infinium® (Illumina Inc., San Diego CA, USA) SNP array that we employed in this study and a list of those genes are available in Geraldes *et al.* (2013), while genotyping is detailed in Porth *et al.* (2013). Genetic distances were calculated with simple Perl scripts. In short, for each pair of trees, we calculated the percentage of shared alleles and calculated the genetic distance as one minus the percentage of shared alleles. Relatedness coefficients were calculated in EigenSoft v4.2 (Patterson *et al.*, 2006). One individual from each pair with relatedness of 0.5 or higher (suggestive of sib relationship) or genetic distance below 0.03% (suggestive of clonal identity) was removed from further analyses. The remaining analyzed population consisted of 334 unrelated individuals.

Genetic structure and trait association analysis

The genetic structure fit was done by performing principal component analysis (PCA) on SNPs (Patterson *et al.*, 2006) with only complete information and satisfying the Hardy–Weinberg expectation (HWE). HWE was tested using 'HWChisq' function implemented in the 'Hardy–Weinberg' R package. PCA was done in R using the 'prcomp' function (Team, 2011) and principal components accounting for 90% of the total variance of the data were retained and subsequently included in the principal component regression analysis needed to select those components affecting the trait in question. This was conducted using the function 'stepwise' implemented in R package 'Rcmdr' with 'backward' selection and Bayesian information criterion 'BIC' as the selection criterion (Pant *et al.*, 2010). The number of principal components (PCs) accounting for 90% of the total variance in the SNP data entering the stepwise selection procedure was 272 and the number of important principal components selected by the backward-stepwise selection differed for each trait and was on average 12 PCs per trait, but ranged from four (average wood density) to 24 (MFA1).

Finally, the selected PCs were included in a regression model along the preselected SNPs (identified for each trait as described later) individually as follows:

$$Y = \mu + S\alpha + \sum_{j=1}^{K} P_j\beta_j + e \qquad \text{Eqn 1}$$

where, Y is vector of measurements, μ is the population mean, α is the SNP effect, $\sum_{j=1}^{K} P_j\beta_j$ represents the effect of selected PCs resulting from the backward-stepwise selection procedure, and e is the residual effect. Association analysis was performed in TASSEL (Bradbury *et al.*, 2007) by the GLM procedure. Corrections for multiple testing were done using a permuted P-value procedure in TASSEL (1000 permutations were run). Cumulative R^2 values were computed using the GLM function implemented in the base package R (Team, 2011), followed by the RsquareAdj function implemented in R package 'vegan' to obtain R^2 for both the reduced and the full model (Peres-Neto *et al.*, 2006).

Two approaches for analyzing SNP–wood chemistry/ultrastructure attributes were tested. First, a two-stage approach (Aulchenko *et al.*, 2007; Pant *et al.*, 2010) in which the initial step, SNP preselection, involved the use of a log-likelihood ratio test for each SNP–trait combination (SNPs = 29 233; traits = 17) performed with the 'anova.lme' function implemented in the 'nlme' R package (Pinheiro & Bates, 2000). SNPs with significant effect ($\alpha \leq 0.01$) were included in the second stage. Secondly, the whole SNP data set was also analyzed using the SimpleM approach for multiple testing corrections (Gao *et al.*, 2008). These results are presented in Fig. S2.

The LDheatmap function implemented in the R package LDheatmap was used to calculate and plot pairwise LD values between all genotyped SNPs, using allelic correlations (r^2), Shin *et al.* (2006).

Results

SNP selection, wood chemistry and ultrastructure traits

The design, composition and performance of the 34K SNP array used in this study, containing 34 131 SNPs in 3543 genes, is detailed in Geraldes *et al.* (2013). Briefly, 99.3% of the SNPs were localized to the 19 *P. trichocarpa* linkage groups that correspond to the 19 pairs of chromosomes. The median distance between SNPs within the 3543 genes was 487 bp, and on average there were 9.5 SNPs per gene region (the transcribed region plus 2 kb of flanking sequence on either side). We used this array to successfully genotype an association population of 334 unrelated individuals collected from wild populations (Geraldes *et al.*, 2013; Porth *et al.*, 2013), representing a geographic range of 44.0–58.6° latitude N. After SNP filtering (see Porth *et al.*, 2013), 29 233 SNPs were available for further analysis.

The population of 334 trees was grown in a common garden in Surrey, BC (49.18°N, 122.85°W), and phenotyped for 17 wood chemistry and ultrastructure traits representing key characteristics affecting pulping, solid wood, and biofuel production (Porth *et al.*, 2013).

Evaluation of phenotype–genotype association approaches

Population and familial stratification are known to affect association studies causing spurious results (Price *et al.*, 2006; Astle & Balding, 2009; Sillanpaa, 2011). We attempted to overcome these effects. Familial structure is *a priori* unlikely to be important in our study population as these are natural accessions of an outbred species, and accessions with evidence of sibship in the association population were eliminated from the analysis (see the Materials and Methods). However, to correct the association model for confounding effects from potential hidden population structure (Patterson *et al.*, 2006) and/or familial structures (Loiselle *et al.*, 1995), we used a subset of SNPs with complete information (i.e. no missing data) and meeting the HWE (N = 9342). Analysis of population structure (using GENELAND; Guillot *et al.*, 2005) and familial structure (Loiselle *et al.*, 1995) indicated that neither had a strong systematic effect (Fig. S1). The population structure analysis discovered two populations with small F_{ST} (0.0121) and, likewise, the coefficients in the kinship matrix showed only small variances. While correction for familial structure (K model) is considered to be an appropriate solution across a wide range of population settings (e.g. Astle & Balding, 2009), this model was not selected in our case (Table S2).

The use of PCA in genome-wide association studies (GWAS) can account for population stratification and LD (Patterson *et al.*, 2006; Price *et al.*, 2006, 2010). Such long-range LD between distant loci can inflate single locus test statistics (quantitative trait nucleotide effects, R^2), whether the result of selection, stratification or genotyping errors (Thomas *et al.*, 2011). As wood traits do not strictly follow the north–south cline (Porth *et al.*, 2013), the use of PCA to control for cryptic genetic background is deemed appropriate to detect genetic associations resulting from phenotypic QTLs in the population. We identified strong LD between SNPs along linkage groups (the extent of LD between the top 20 SNPs in the simple model for galactose is shown in Fig. S3) and thus we fitted LD across the whole genome. To account for the cryptic genetic background, we employed backward-stepwise regression (i.e. PCA-based model that selects PCs via BIC), and PCs with eigenvalues explaining 90% of the total variation in genomic data were included in the analysis (Pant *et al.*, 2010). Examination of qq-plots comparing alternative methods applied to our data showed that, for most traits, the PCA-BIC method outperformed the other methods (Fig. S2). When we compared R^2 values generated by two PCA-based alternative methods, we found that the backward-stepwise BIC-based PC selection approach effectively corrected for the impact of LD on the SNP effects (Fig. S2). The lack of significant associations in the simple (no adjustment) model for most traits, however, indicates lower analysis power, which could be caused by an unfavorable balance between the number of individuals and the number of SNPs or by the genetic nature of the studied population. However, it is most likely that the power of the analysis is impacted most by the modest size of the association population (Fig. S4).

For determining statistically significant SNP–wood phenotype associations, we first used a two-stage approach in which the initial step, SNP preselection, involved the use of a log-likelihood ratio test for each SNP–trait combination. SNPs with significant effect ($\alpha \leq 0.01$) were included in the second stage. Including only a set of meaningful SNPs in the subsequent analysis greatly reduces the effect of multiple testing corrections when testing a large initial number of SNPs for associations (Aulchenko *et al.*, 2007; Macciotta *et al.*, 2009; Pant *et al.*, 2010). Secondly, for comparison, the whole SNP data set was analyzed using a modified Bonferroni correction that takes LD into account and is therefore based on the effective number of tests (Gao *et al.*, 2008). In summary, the power of the association analysis was increased when we preselected SNPs using a log-likelihood ratio test to include only SNPs with potential associations with the studied traits (Aulchenko *et al.*, 2007; Macciotta *et al.*, 2009; Pant *et al.*, 2010), while the backward-stepwise BIC-based PC selection approach best corrected for the impact of LD on SNP effects.

Phenotype–genotype associations

The number of SNPs preselected by the log-likelihood ratio test at $\alpha = 0.01$ was trait-specific and encompassed on average 181 SNPs per trait, but ranged between 71 and 662 SNPs for holocellulose and soluble lignin content, respectively. Employing the PCA-BIC/SNP preselection pipeline described earlier, we identified 141 SNP–trait associations that were significant at $\alpha \leq 0.05$ and are distributed among the 19 *Populus* chromosomes (Table 1, Fig. 1). Our analysis identified SNP associations with 16 of the 17 secondary cell wall and wood ultrastructure traits examined, failing to find significant associations for galactose content. The percentage of phenotypic trait variance attributed to an individual SNP ranged from 2.9 to 6.9%, based on R^2 values (Tables 1, S3). We also fitted a model where all significantly associated SNPs were included as explanatory variables for the respective traits to estimate the cumulative phenotypic variance (cumulative R^2) explained by the genetic effects of all SNPs associated with a trait, calculated according to Ingvarsson *et al.* (2008). The cumulative phenotypic variance explained by all significant SNPs ranged between 5 and 26% (Table 1). While estimates roughly reflect the number of SNP associations for the traits, the portion of the total phenotypic variance explained by all significant SNPs is particularly dependent on the magnitudes of the allelic effects (Table 1) and also on the existing LD between these SNPs. The majority of associations were found in markers with small minor allele frequency ($0.05 < \text{MAF} < 0.2$; Fig. 2a). While no correlation between MAF and R^2 was detected (Fig. 2b), large variances in allelic effects were found for low-MAF markers (Fig. 2c) related to the variance explained by an SNP when only additive gene action is present (Falconer, 1981).

We previously determined the narrow-sense heritability (h^2) for all 16 wood traits associated with SNP variants, based on the realized kinship between all individuals in the field trial (Porth *et al.*, 2013). Most traits showed moderate to high heritability (h^2 between 0.4 and 0.7). Knowing both the genetic effects of SNPs associated with a phenotypic trait and h^2 of that individual trait allows the total genetic variance explained by the associated SNPs to be determined (Gonzalez-Martinez *et al.*, 2007). Thus, we were able to partition the genetic variance explained by each significant SNP (Table 1).

We detected from two (xylose content) to 15 (holocellulose content) SNP associations for 16 traits (Table 1). The allelic mode of gene action found for these polymorphisms included additive, dominance, overdominance and underdominance effects (Table 1). The most common modes of allelic effects were dominance and over/underdominance. The majority of the 141 SNPs with significant associations were located in noncoding regions (80%), while 28 SNPs were found within coding regions, of which half created nonsynonymous polymorphisms. However, many of the SNPs in noncoding regions are expected to be in LD with coding-region SNPs based on the haplotype tagging strategy employed in SNP selection for the array.

Wood ultrastructural traits yielded, on average, more significant associations per trait than wood chemical traits (Table 1). For several industrially important cell wall traits (e.g. glucose content, fiber length, cell wall crystallinity), single SNPs (Table 2) explained a relatively large portion of the variation. For example, SNP scaffold_11_18127986 within POPTR_0011s16200 (*FAD-binding domain-containing protein* gene), associated with total lignin, had the highest R^2 value in the data set (6.9%). The largest values for SNP effects on glucose and syringyl lignin content were 5.7% (POPTR_0018s11290, *PtiCESA7-B*) and 5.3% (POPTR_0018s12720, *ARF-GAP domain 11* gene), respectively. We identified 21 genes with more than one SNP significantly associated with variation in the same trait (12 traits total; Tables S3, S4). This is illustrated by the allelic effects of 14 significant wood density SNP associations in Fig. 3. For SNPs that are in high linkage, phenotypic effects showed similar patterns (Fig. 3), indicating haplotype blocks that underlie variation in the identified traits. Therefore, it is impossible to determine if one or all have a functional effect.

Although the data do not indicate which SNPs may be causal, it is interesting to note that some of these SNPs represent amino acid replacement (nonsynonymous) polymorphisms, such as the SNPs in *FRA8* (the *Populus* ortholog of *Arabidopsis FRAGILE FIBER 8*), which is associated with fiber length (g. 4), and *CRLK42* (POPTR_0012s08890 annotated as *CYSTEINE-RICH*

Table 1 Summary of single nucleotide polymorphism (SNP)–trait associations and SNP allele effects from a genome-wide association analysis of *Populus trichocarpa*

						SNP allele effect[f]			
Trait no.	Trait[a]	*N*[b]	Phenotypic variance (%)[c]	Genetic variance (%)[d]	Phenotypic variance (cumulative) (%)[e]	D	A	U/OD	No H2
1	Alpha cellulose	12	3.1–4.0	7.5–9.6	8.6	0	4	6	2
2	Arabinose	3	3.5–3.7	5.1–5.4	4.5	0	0	3	0
3	Average wood density	14	3.4–6.5	6.1–11.6	22.2	4	5	4	1
4	Crystallinity	8	3.8–4.9	9.5–12.3	16.9	5	2	1	0
5	Fiber LW	11	3.6–5.1	15.0–21.3	16.5	3	4	4	0
6	Galactose	0	–	–	–	–	–	–	–
7	Glucose	7	3.5–5.7	7.7–12.5	10.1	1	0	6	0
8	Hemicellulose	9	3.1–4.4	8.9–12.7	12.6	2	0	6	1
9	Holocellulose	15	2.9–5.9	n.a.	17.8	5	2	7	1
10	Insoluble lignin	8	4.4–5.3	6.7–8.1	7.3	7	0	1	0
11	Mannose	12	4.2–5.6	10.9–14.5	22.5	3	3	6	0
12	MFA1	7	3.2–3.7	7.3–8.4	5.5	2	3	2	0
13	MFA2	12	3.8–5.4	9.3–13.3	21.8	5	3	4	0
14	Soluble lignin	3	4.1–5.8	4.2–6.0	7.7	0	1	2	0
15	Syringyl lignin	12	4.0–5.3	10.3–13.7	26.2	6	3	3	0
16	Total lignin	6	3.9–6.9	6.0–10.6	7.9	3	1	2	0
17	Xylose	2	4.1–4.4	6.0–6.4	5.2	1	0	1	0
	Total	141				47	31	58	5

[a] Trait definitions: alpha cellulose, percentage of alpha cellulose in extract-free dry wood; arabinose, percentage of arabinose sugar in extract-free dry wood; average wood density, wood density averaged from pith to bark (in kg m^{-3}); crystallinity, percentage of cell wall crystallinity; fiber LW, cellulose fiber length weighted (in mm); galactose, percentage of galactose sugar in extract-free dry wood; glucose, percentage of glucose sugar in extract-free dry wood; hemicellulose, percentage of hemicelluloses in extract-free dry wood; holocellulose, percentage of holocellulose in extract-free dry wood; insoluble lignin, percentage of Klason lignin in extract-free dry wood; mannose, percentage of mannose sugar in extract-free dry wood; MFA1, microfibril angle measured at recent growth ring (in degree); MFA2, microfibril angle measured at first growth ring (in degrees); soluble lignin, percentage of acid soluble lignin in extract-free dry wood; syringyl lignin, percentage of syringyl monomers in extract-free dry wood; total lignin, percentage of total lignin in extract-free dry wood; xylose, percentage of xylose sugar in extract-free dry wood.

[b] Number of significant associations at a given alpha level ($\alpha \leq 0.05$).

[c] Range of percentage trait variance explained by genetic effects of associated SNPs (output from TASSEL).

[d] Range of percentage heritable trait variance explained by genetic effects of associated SNPs; genetic variance is calculated on the basis of previously determined heritability estimates (Porth *et al.*, 2013).

[e] Phenotypic variance (cumulative R^2) explained by the genetic effects of all associated SNPs calculated according to Ingvarsson *et al.* (2008) using the GLM procedure.

[f] Number of SNPs in each of four classes: D, dominant (heterozygote similar to one homozygote); A, additive allele action (heterozygote intermediate); U/OD, under- or overdominant (heterozygote outside the range of the homozygotes); No H2, one homozygous genotype class was missing, thus no gene mode was estimated; estimation of mode of gene action according to Wegrzyn *et al.* (2010).

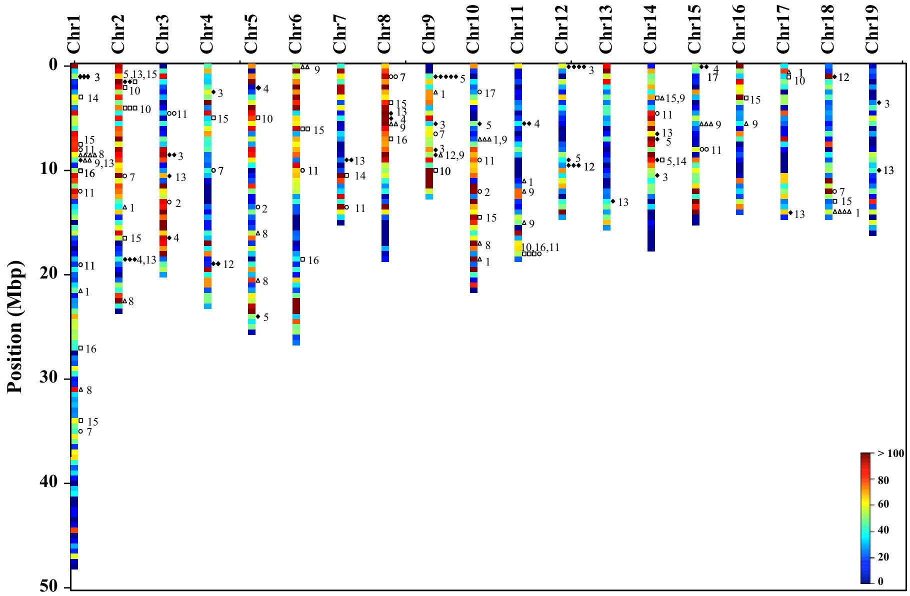


Fig. 1 Genome-wide depiction of significant single nucleotide polymorphism (SNP)–trait associations in *Populus trichocarpa* for 16 phenotypes. The locations of 141 associations significant at $\alpha \leq 0.05$ along chromosomes 1–19 of the *P. trichocarpa* genome are shown. Open circles, associations with cell wall sugars (arabinose, glucose, mannose, and xylose); open squares, associations with lignin (insoluble, soluble, syringyl, and total lignin); open triangles, associations with polysaccharides (alpha cellulose, hemicelluloses, and holocellulose); and black diamonds, associations with ultrastructure (average density, crystallinity, fiber length, MFA1, and MFA2) and individual trait label (number) according to order in Table 1. SNP density of the 34K SNP genotyping array in 0.5 Mbp windows across the genome follows the color scheme at the bottom right of the figure.

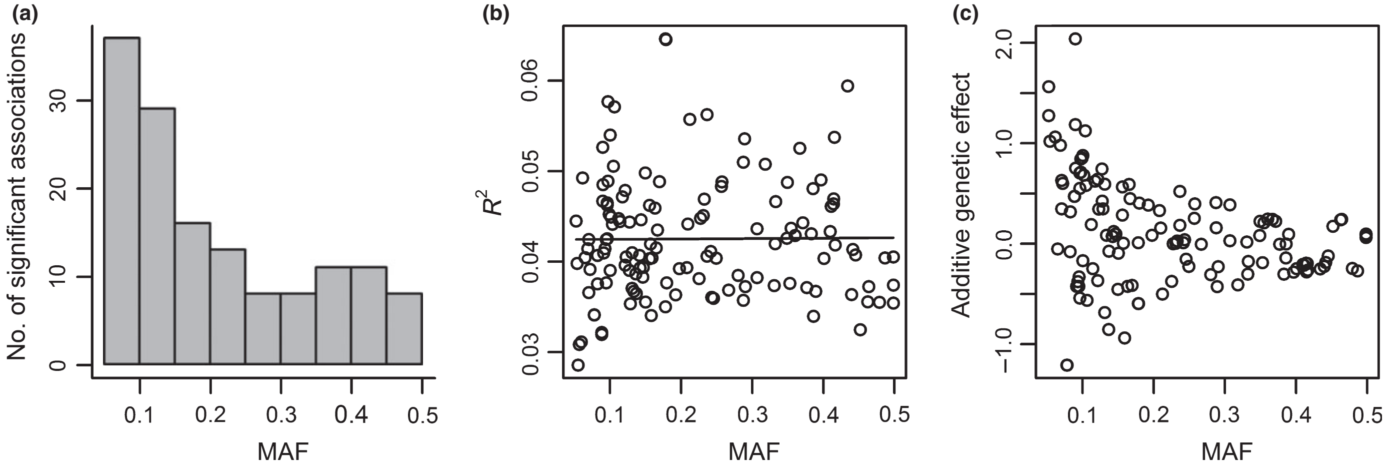


Fig. 2 The effect of minor allele frequency (MAF) on the genetic and the allelic effects in the significant single nucleotide polymorphism (SNP)–trait associations in *Populus trichocarpa*. (a) Number of significant SNP associations vs MAF; (b) R^2 vs MAF; (c) additive genetic effect vs MAF.

RLK RECEPTOR-LIKE PROTEIN KINASE 42), associated with MFA1 (Fig. S5).

The significant trait-associated SNPs were derived from multiple alternate functional and expression categories (see the FamilyPath column in Table S3). For example, 21 SNPs found within genes encoding 18 transcription factors were associated with 10 different traits, such as alpha cellulose or mannose content (Tables S5,S3, Fig. S6), and 17 SNPs found within 11 genes related to cell wall metabolism were identified for eight different traits (Tables S3, S6). Some of these genes encode structural proteins (e.g. arabinogalactan protein), while others function in processes such as monosaccharide activation and interconversion (nucleotide–sugar interconversion enzyme activity), polysaccharide synthesis (cellulose and hemicellulose synthesis), cell wall reassembly (glycoside hydrolase) or glycosyl transferase activities. A large set of associations (60 SNPs within 39 genes associated with 15 different traits) were identified in genes included in the 34K SNP genotyping array based on transcript profiling or expression patterns rather than protein functional information (Geraldes *et al.*, 2013; Table S7).

Genetic associations with phenotypic variation in cell wall carbohydrate content

We identified 48 SNPs associated with cell wall carbohydrate traits (Table 1). Among these, the synonymous SNP, scaffold_18_12097363, located within *PtiCESA7-B*, a *P. trichocarpa* ortholog (POPTR_0018s11290) of *Arabidopsis CESA7* (*IRREGULAR XYLEM 3*) encoding a cellulose synthase subunit implicated in cellulose biosynthesis in the secondary cell wall (M. Kumar *et al.*, 2009), showed genetic association with glucose content. The genetic effect of this SNP explained 5.7% of the total phenotypic variance and 12.5% of the total genetic variance when considering heritability. This polymorphism showed an underdominance effect where the heterozygous genotype had lower glucose content than either of the two homozygous genotypes (Fig. 5a).

In other examples, two SNPs associated with relative xylose content (Fig. 5b): one intronic SNP, scaffold_10_2623857 within POPTR_0010s02280 (annotated as similar to *FASS*, protein phosphatase type 2B regulator), explained 4.1% of the total phenotypic variance, and one SNP, scaffold_15_912654, located within the 5′ UTR flanking region of POPTR_0015s01480 (annotated as similar to IQ-domain 21; IQD21 calmodulin-binding protein), explained 4.4% of the total phenotypic variance for this trait. Based on previous results on xylose trait heritability in this population (Porth *et al.*, 2013), these two SNPs account for 6.0–6.4% of the total genetic variance in xylose content (Table 1). Examples of associations with relative holocellulose content (the combination of both alpha cellulose and total hemicelluloses) are two SNPs within POPTR_0001s11720, which encodes a Rab GTPase similar to the *Arabidopsis ARA-5* (Figs S7, S8).

Finally, four SNPs within three different intronic regions of POPTR_0018s13970 (encoding an ENTH domain protein) were associated with alpha cellulose content. Although separated by up to 3 kb, all four SNPs were in complete LD (Fig. S9). Summaries of SNP associations (boxplots) with variation in individual cell wall monomeric sugars are presented in Fig. S10.

Genetic associations with phenotypic variation in lignin content

We identified 48 SNPs associated with lignin traits (Table 1). The boxplots for all lignin traits including the summaries of significant SNP associations are shown in Fig. S11. The SNP scaffold_11_18127986 in POPTR_0011s16200 (annotated as *FAD-binding domain-containing protein*) is associated with variation in insoluble lignin content as well as the amount of total lignin (Fig. 6), traits that are highly correlated (Porth *et al.*, 2013). A second SNP in the 5′-upstream region of the gene and in linkage showed significant association with total lignin only, and exhibited a similar dominance allelic affect. Another example of dominance for associations with insoluble lignin involved three intronic SNPs within POPTR_0002s06080 (*Populus MAP KINASE20-2*; *PtMPK20-2*; Fig. S12).

An intronic SNP at position scaffold_17_1039469 within POPTR_0017s01510 (*PttMAP20*) encoding a microtubule

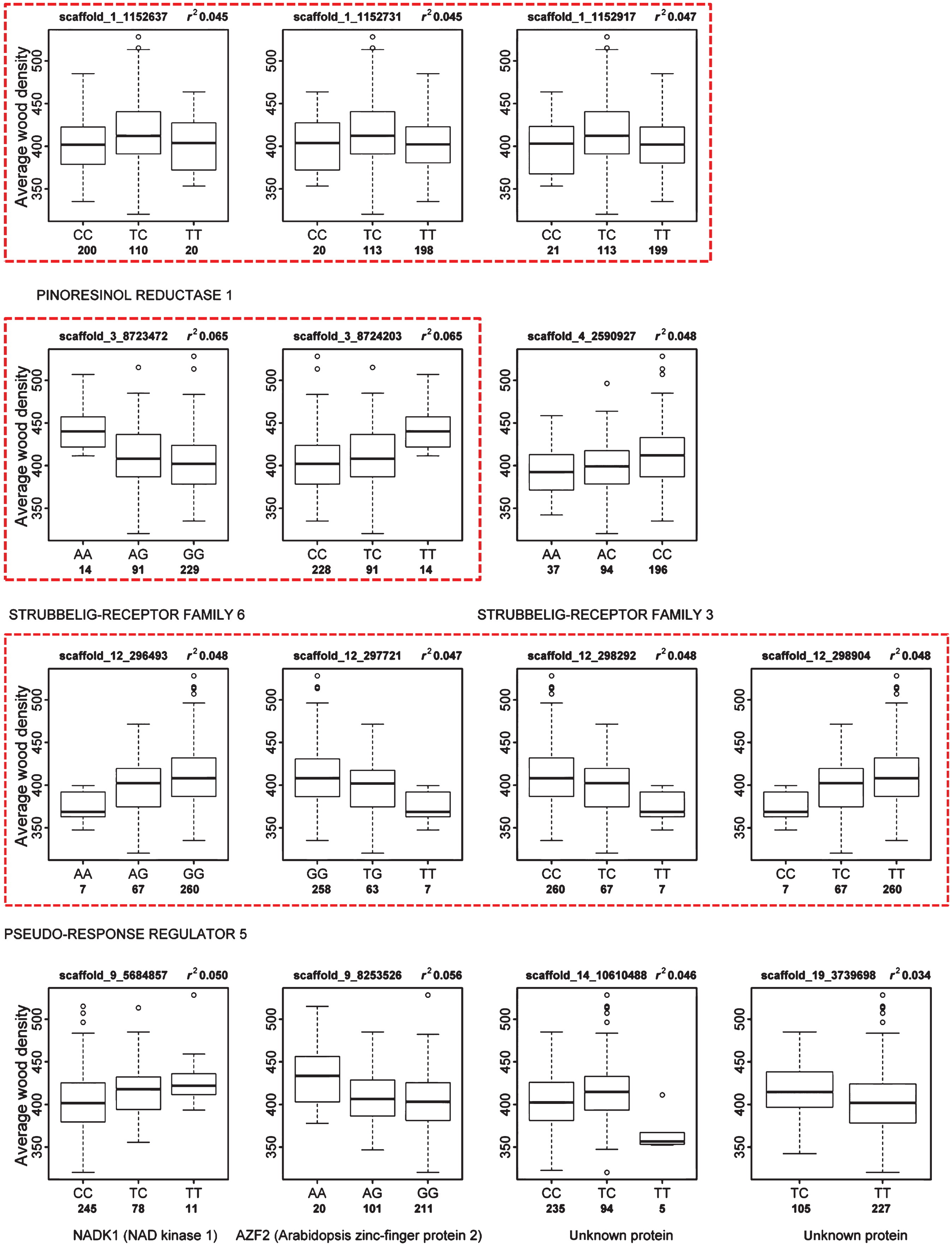
Table 2 Single nucleotide polymorphisms (SNPs) with the largest effect on phenotypic variance (R^2) in *Populus trichocarpa*

Trait	SNP position	R^2	SNP location, feature	Gene model	Best Arabidopsis hit	Annotated function
Total lignin	scaffold_11_18127986	0.069	Upstream, intergenic	POPTR_0011s16200	AT1G30760	FAD-binding domain-containing protein
Average wood density	scaffold_3_8723472	0.065	Intron	POPTR_0003s07130	AT1G53730	STRUBBELIG-RECEPTOR FAMILY 6 (SRF6)
Average wood density	scaffold_3_8724203	0.065	Intron	POPTR_0003s07130	AT1G53730	STRUBBELIG-RECEPTOR FAMILY 6 (SRF6)
Holocellulose	scaffold_1_9088450	0.059	Upstream, intergenic	POPTR_0001s11720	AT1G02130	ARABIDOPSIS RAS 5 (ARA-5)
Soluble lignin	scaffold_1_3044269	0.058	Upstream, intergenic	POPTR_0001s03760	AT1G55570	SKU5 Similar 12 (sks12)
Glucose	scaffold_18_12097363	0.057	CDS, S	POPTR_0018s11290	AT5G17420	PtiCESA7-B IRX3 (IRREGULAR XYLEM 3)
Mannose	scaffold_11_18475792	0.056	Intron	POPTR_0011s16690	AT1G30580	GTP binding
Average wood density	scaffold_9_8253526	0.056	Downstream, intergenic	POPTR_0009s09250	AT3G19580	ARABIDOPSIS ZINC-FINGER PROTEIN 2 (AZF2)
Mannose	scaffold_7_13909650	0.054	Downstream, intergenic	POPTR_0007s13910	AT2G18060	VASCULAR RELATED NAC-DOMAIN PROTEIN 1 (VND1)
Mannose	scaffold_1_8367760	0.054	Downstream, intergenic	POPTR_0001s10750	AT1G64480	CALCINEURIN B-LIKE PROTEIN 8 (CBL8)
MFA2	scaffold_19_10226694	0.054	Intron	POPTR_0019s08690	No Arabidopsis blast hit	
Insoluble lignin	scaffold_2_4071049	0.053	Intron	POPTR_0002s06080	AT2G42880	ATMPK20
Insoluble lignin	scaffold_2_4071255	0.053	Intron	POPTR_0002s06080	AT2G42880	ATMPK20
Insoluble lignin	scaffold_2_4074112	0.053	Intron	POPTR_0002s06080	AT2G42880	ATMPK20
Syringyl lignin	scaffold_18_13238386	0.053	CDS, NS	POPTR_0018s12720	AT3G07490	ARF-GAP domain 11 (AGD11)
Fiber LW	scaffold_9_1380717	0.051	CDS, NS	POPTR_0009s01200	AT2G28110	FRAGILE FIBER 8 (FRA8)
Insoluble lignin	scaffold_17_1039469	0.051	Intron	POPTR_0017s01510	AT5G37478	Unknown protein
Insoluble lignin	scaffold_2_2323118	0.051	Downstream, intergenic	POPTR_0002s03730	AT4G28540	CASEIN KINASE I-LIKE 6 (CKL6)
Average wood density	scaffold_9_5684857	0.050	Intron	POPTR_0009s05620	AT3G21070	NAD KINASE 1 (NADK1)
Total lignin	scaffold_11_18128034	0.049	Upstream, intergenic	POPTR_0011s16200	AT1G30760	FAD-binding domain-containing protein
Crystallinity	scaffold_11_5726524	0.049	Intron	POPTR_0011s05740	AT4G28500	ARABIDOPSIS NAC DOMAIN CONTAINING PROTEIN 73 (ANAC073)
Fiber LW	scaffold_9_1379264	0.049	Downstream, intergenic	POPTR_0009s01200	AT2G28110	FRAGILE FIBER 8 (FRA8)
Mannose	scaffold_15_8033813	0.049	Upstream, intergenic	POPTR_0015s06980	AT5G60970	TEOSINTE BRANCHED1, CYCLOIDEA AND PCF TRANSCRIPTION FACTOR 5 (TCP5)
Insoluble lignin	scaffold_5_5353766	0.049	CDS, S	POPTR_0005s07810	AT4G39840	Unknown protein
Mannose	scaffold_14_4655035	0.049	Upstream, intergenic	POPTR_0014s06150	AT1G02030	Zinc finger (C2H2 type) family protein
Syringyl lignin	scaffold_6_6390311	0.049	Upstream, intergenic	POPTR_0006s08720	AT5G02230	Haloacid dehalogenase-like hydrolase family protein

CDS, coding sequence; NS, nonsynonymous substitution; S, synonymous substitution.

binding protein whose expression is strongly correlated with secondary wall formation in *Populus* (Rajangam *et al.*, 2008) was associated ($R^2 = 5.1\%$) with insoluble lignin (Fig. 5c). The mode of the allelic gene action is dominant, with T as the dominant allele (lower insoluble lignin content). The genetic variance for insoluble lignin that was explained by this SNP amounted to 7.8%. A final example of a lignin trait-associated polymorphism is SNP scaffold_14_9134462 within POPTR_0014s12380, whose closest *Arabidopsis* homologue is the *AUD1* gene encoding UDP-glucuronate decarboxylase involved in xylose metabolism. This SNP explained < 4% of the total genetic variance in

Fig. 3 Boxplot diagrams depicting the genetic effects of all single nucleotide polymorphisms (SNPs) with significant associations to average wood density in *Populus trichocarpa*. The box in each box plot shows the lower quartile, and the median and upper quartile values, and the whiskers show the range of the phenotypic variation in the population; red dashed-lined frames indicate SNPs in linkage disequilibrium (LD). Wood density units in are in kg m^{-3}. PRR1 (PINORESINOL REDUCTASE 1), scaffold_1_1152637, scaffold_1_1152731 and scaffold_1_1152917; SRF6 (STRUBBELIG-RECEPTOR FAMILY 6), scaffold_3_8723472, scaffold_3_8724203; SRF3 (STRUBBELIG-RECEPTOR FAMILY 3), scaffold_4_2590927; NADK1 (NAD KINASE 1), scaffold_9_5684857; AZF2 (ARABIDOPSIS ZINC-FINGER PROTEIN 2), scaffold_9_8253526; PSEUDO-RESPONSE REGULATOR 5, scaffold_12_296493, scaffold_12_297721, scaffold_12_298292 and scaffold_12_298904; unknown protein, scaffold_14_10610488; unknown protein, scaffold_19_3739698.


scaffold_1_1152637 r^2 0.045
scaffold_1_1152731 r^2 0.045
scaffold_1_1152917 r^2 0.047
Average wood density
350
400
450
500
CC 200
TC 110
TT 20
CC 20
TC 113
TT 198
CC 21
TC 113
TT 199
PINORESINOL REDUCTASE 1
scaffold_3_8723472 r^2 0.065
scaffold_3_8724203 r^2 0.065
scaffold_4_2590927 r^2 0.048
AA 14
AG 91
GG 229
CC 228
TC 91
TT 14
AA 37
AC 94
CC 196
STRUBBELIG-RECEPTOR FAMILY 6
STRUBBELIG-RECEPTOR FAMILY 3
scaffold_12_296493 r^2 0.048
scaffold_12_297721 r^2 0.047
scaffold_12_298292 r^2 0.048
scaffold_12_298904 r^2 0.048
AA 7
AG 67
GG 260
GG 258
TG 63
TT 7
CC 260
TC 67
TT 7
CC 7
TC 67
TT 260
PSEUDO-RESPONSE REGULATOR 5
scaffold_9_5684857 r^2 0.050
scaffold_9_8253526 r^2 0.056
scaffold_14_10610488 r^2 0.046
scaffold_19_3739698 r^2 0.034
CC 245
TC 78
TT 11
AA 20
AG 101
GG 211
CC 235
TC 94
TT 5
TC 105
TT 227
NADK1 (NAD kinase 1)
AZF2 (Arabidopsis zinc-finger protein 2)
Unknown protein
Unknown protein

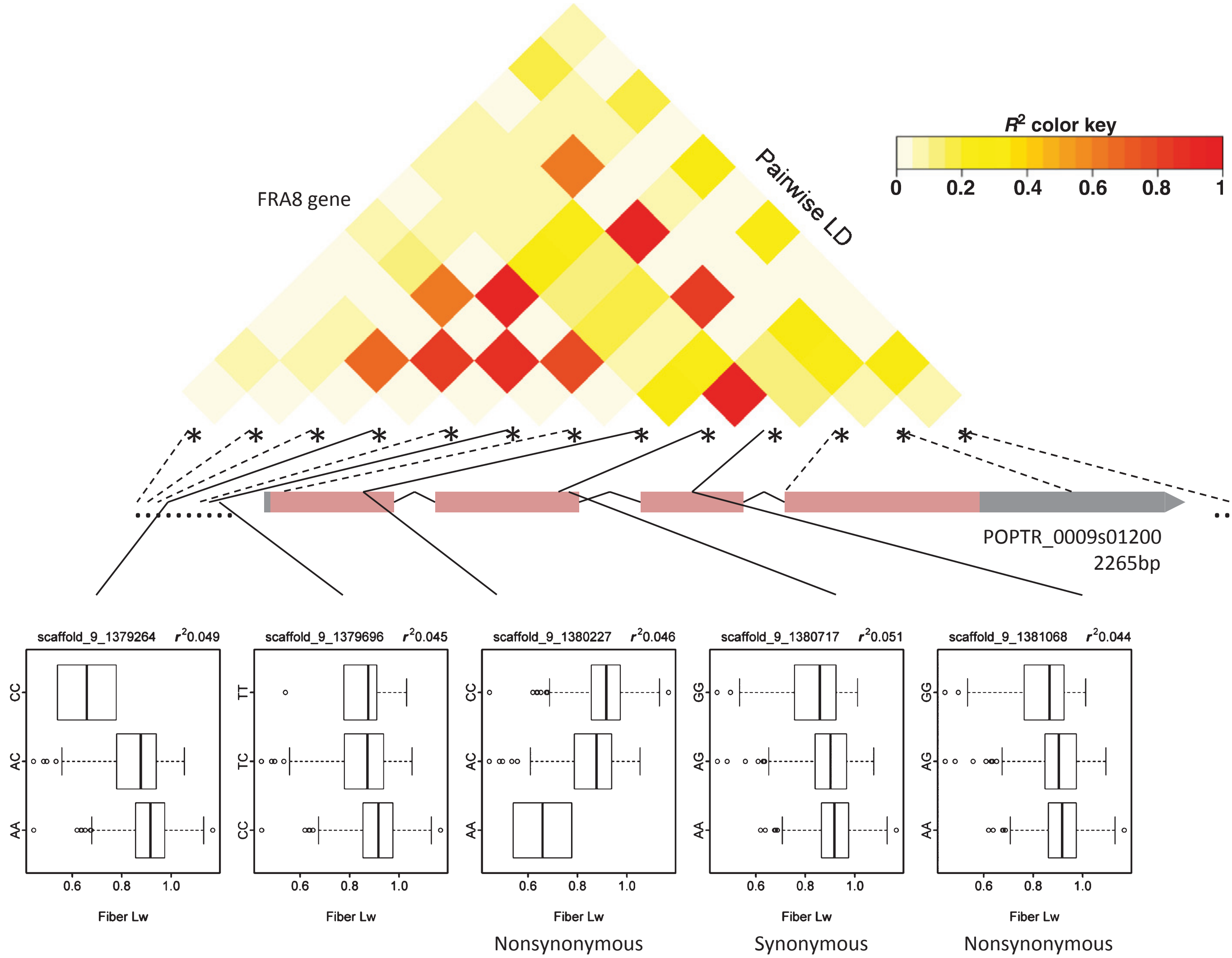


Fig. 4 *FRA8* gene structure, linkage of genotyped single nucleotide polymorphisms (SNPs), and genetic effects of five SNPs associated with fiber length (Fiber Lw) in *Populus trichocarpa*. Gene structure is represented schematically as exon (boxes)-intron (lines)-UTR (dots) patterns; the solid lines connect to the five associations ($\alpha \leq 0.05$, explaining 4.4–5.1% of fiber length (mm) variation) for which boxplot diagrams depict their genetic effects. Dashed lines indicate other genotyped SNPs with no significant association with fiber length. The box in each box plot shows the lower quartile, and the median and upper quartile values, and the whiskers show the range of the phenotypic variation in the population (units in mm). The linkage disequilibrium (LD) structure (r^2) is color-coded to show the extent of LD across the genotyped SNPs with the candidate gene.

soluble lignin with an estimated over/underdominance effect (Fig. S13).

Genetic associations with wood ultrastructure variation

We also determined variation in five wood ultrastructure traits (Table 1, Figs 1, S14). The individually associated SNPs explained a relatively high portion of the genetic variance of these traits. The genome-wide distribution of SNP associations with fiber length is shown in Fig. S15. For fiber length, five significant SNP associations, including three SNPs in coding regions (two nonsynonymous substitutions), were detected within POPTR_0009s01200 (*FRA8*) (Fig. 4), providing strong evidence that genetic variation at this gene underlies variation in fiber length. The strongest genetic SNP effect was found at position scaffold_9_1380717 (a synonymous allelic variant in *FRA8*), explaining 5.1% of the total phenotypic variance and 21% of the genetic variance in fiber length. Interestingly, while three out of these five SNPs followed an additive gene model, the remaining two followed either an over/underdominance or complete dominance model (Fig. 4).

For genetic associations with average wood density, we identified three SNPs in LD within POPTR_0001s01540 (*PRR1*) (Fig. S16). The SNP at position scaffold_1_1152917 in the 5′-gene region explained 4.7% of the phenotypic and 8% of the genetic variance of wood density. The majority of the identified SNP associations with wood density followed an additive gene model (Fig. 3).

Three significant SNP associations to microfibril angle (MFA1, recent growth ring) were found within POPTR_0012s08890, homologous to *Arabidopsis CRLK42* (Fig. S5), including one nonsynonymous nucleotide replacement polymorphism (scaffold_12_9951792). All three SNPs were in strong

LD, with an r^2 of *c.* 0.7; however, SNP scaffold_12_9949191 exhibited the strongest association and is located within the 3′ gene region explaining 8% of the genetic variance for MFA1. All three SNPs follow the additive gene model; however, individuals homozygous for the favorable alleles for low MFA (advantageous trait) were also extremely rare in the population (Fig. S14). For cell wall crystallinity, we identified two significant SNP associations within an intronic region of POPTR_0011s05740 annotated as *ANAC073* (Fig. S17).

Discussion

We investigated a comprehensive array of wood traits across the natural range of *P. trichocarpa* to identify potential genetic markers for important wood properties for industrial applications. The present work reports the identification of a set of candidate allelic variants that could be exploited to alter cell wall traits for improved pulping, solid wood properties and cellulosic ethanol production (Fig. 1). We employed a broadly selected suite of genes to disclose SNP associations with the key industrial traits. We previously showed that the improvement of *P. trichocarpa* by simultaneous selection of desirable industrial wood trait combinations for feedstock improvement (e.g. high cellulose and low lignin) should be possible given the genetic correlations between these traits in our association population (Porth *et al.*, 2013).

Genetic associations can be obscured by confounding factors (population and familial structure) for which the association models should be corrected. Tracy–Widom statistics are often used to fit population structure in association studies through the selection of the most important PCs (Patterson *et al.*, 2006; Price *et al.*, 2006); however, Shriner (2011) found that this approach may yield the inclusion of additional significant PCs as a result of the effect of admixture, thus adding noise and ultimately affecting the power of the analysis. Moreover, the selected PCs should be related to phenotypes so that the added noise should not affect the employed association model (Novembre & Stephens, 2008; Zhu & Yu, 2009; Peloso & Lunetta, 2011). The present analysis of population structure did not reveal strong systematic stratification (Fig. S1), which is in agreement with a previous study (Slavov *et al.*, 2012). However, Slavov *et al.* (2012) suggested that the unique population structure of black cottonwood along with local LD should be considered in association genetics analyses. Therefore, we focused on fitting the genetic background, including the long-distance LD through the proper selection of PCs that truly relate to the phenotype under investigation (Pant *et al.*, 2010). This was done using the back propagation algorithm proposed by Setakis *et al.* (2006).

Our results identified multiple significant associations individually explaining a small portion of the phenotypic variance. This finding is in agreement with the polygenic character of wood attributes (Gonzalez-Martinez *et al.*, 2007). The sample size available for the present study ($N=334$) resulted in associations individually explaining *c.* 4–6% of the total phenotypic variance (Table 1). These values are higher than those reported in other studies on poplar species with similar traits (Wegrzyn *et al.*, 2010). These discrepancies can be caused by the 'Beavis effect' (Allison *et al.*, 2002; Xu, 2003), resulting in considerably inflated allelic effects for smaller sample sizes (Ingvarsson *et al.*, 2008). In principle, association genetics needs to apply very stringent statistical thresholds. While a large portion of the genetic variation has to be captured for plant breeding, the high degree of statistical significance expected in association studies may not be appropriate in the application phase (tree breeding). While we note here that, generally, the power analysis is not connected to the multiple testing correction (Teyssèdre *et al.*, 2012), our analysis (Fig. S4) clearly revealed that the potential limitation of population size mostly involves the ability to detect small effect QTLs when there is little LD between a SNP marker and a QTL. The majority of associations were found in markers with smaller MAF (< 0.2; Fig. 2a). However, MAF had no significant impact on R^2 (Fig. 2b), suggesting that our MAF cutoff (5%) did not yield associations with artificially inflated R^2 estimates at lower MAF. Furthermore, the large variances in the additive genetic effects for low-MAF markers (Fig. 2c) suggest that such polymorphisms (rarer variants) may indeed explain a substantial portion of the phenotype variation (Dickson *et al.*, 2010; Eichler *et al.*, 2010). We also observed that most modes of gene action were consistent with over/underdominance (58 cases). Overdominance allelic effects suggest that heterozygous individuals perform better than the respective homozygous individuals in terms of greater fitness. In a population genetics context, such individuals help to maintain the genetic variability in a population (through high polymorphism) and are likely to have a natural advantage (Slatkin & Muirhead, 1999) by the combination of both alleles (overdominant selection).

Knowledge of the genetic control for these wood traits permitted further inferences to be made about the importance of trait-associated SNPs (Porth *et al.*, 2013). That is, additive models that combine the genetic effects from multiple SNPs were used to explain a substantial portion of the trait's heritability, such as cell wall crystallinity, MFA, syringyl lignin, total lignin, average wood density and mannose content. At the individual SNP level, the highest phenotypic variance that was accounted for by an individual allelic variant was *c.* 7% for total lignin, with a single SNP in POPTR_0011s16200, encoding an *FAD-binding domain-containing protein* (Fig. 6). Such distinct polymorphisms, as they are also associated with highly heritable traits, promise the highest potential for selective tree breeding, provided their future validation in unrelated approaches such as other genetically distinct and larger association populations. Interestingly, with one exception (the *Populus CESA7-B* gene), none of our 141 SNP–trait associations were in genes that were in common with those identified in a previous study in this species, which analyzed 846 SNPs in 40 candidate genes (Wegrzyn *et al.*, 2010), suggesting our genome-wide analysis has uncovered a richer representation of the genetic architecture of wood trait variation.

Polymorphisms with high potential to improve wood properties for industrial applications

Cellulose content and properties Alpha cellulose accounts, on average, for *c.* 45% of the DW of *Populus* wood. For the production of bioethanol from cellulosic biomass, a breeding target is to generate trees that exhibit high productivity and can accumulate higher relative amounts of cellulose. In our study, several SNP associations with cellulose and glucose content were detected. Importantly, several of these allelic variants explained a relatively high portion of the observed phenotypic variance (4–6%) and were located within *Populus* genes that are strong candidates for involvement in secondary cell wall formation based on known functions and/or expression patterns (e.g. *ARA-5*, (Persson *et al.*, 2005); *ATOPT3*, (Ko & Han, 2004); *ENTH domain protein*; *Populus CESA7-B* (Taylor *et al.*, 2003); *ATMPK3*, (Ko & Han, 2004); *REVOLUTA*; Tables S4,S7). Wegrzyn *et al.* (2010) previously reported significant association of an SNP in *CESA7-B* (also referred to as *PtCesA2.2*; see M. Kumar *et al.*, 2009), with variation in C6 sugars in the wood of a *P. trichocarpa* association population, yet only a small portion of the phenotypic variance (*c.* 1.6%) was explained by this nonsynonymous polymorphism. The synonymous SNP identified in our study explained almost 6% of the phenotypic variance in glucose content (Fig. 5a), consistent with the central role played by CESA7, together with CESA4 and CESA8, as subunits of the CESA complex required for proper secondary cell wall cellulose synthesis (M. Kumar *et al.*, 2009). *Populus* genotypes with high glucose content homozygous for the favorable allele (G) were rare in the population we sampled (Fig. 5a), suggesting a recent mutation in the population, still at low frequency, or alternatively, a possible deleterious fitness effect in nature. However, no negative effect on growth rate and biomass accumulation was observed in these individuals in an association trial (A. D. McKown *et al.*, unpublished), suggesting a potential for deployment in managed stands.

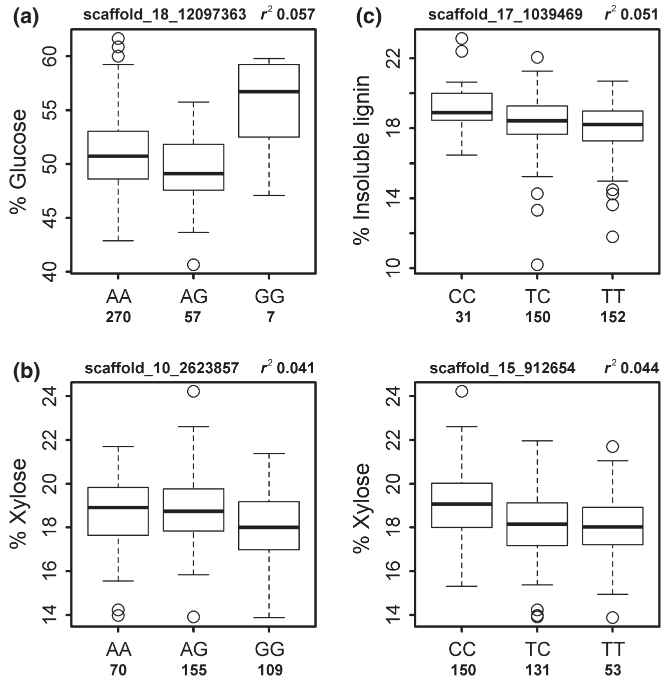


Fig. 5 Boxplot diagrams depicting the genetic effects of single nucleotide polymorphisms (SNPs) with significant associations to glucose, xylose, and insoluble lignin content in *Populus trichocarpa*. (a) Boxplot diagram depicting the genetic effect of an SNP detected in PtiCESA7-B with a significant association to relative glucose content (percentage in extract-free dry wood); (b) Boxplot diagrams depicting the genetic effects of two SNPs in *FASS1* (scaffold_10_2623857) and *IQD21* (scaffold_15_912654) with significant associations to xylose content (percentage in extract-free dry wood); (c) Boxplot diagram depicting the genetic effect of an SNP detected within *PttMAP20* with a significant association to insoluble lignin content (percentage in extract-free dry wood). The box in the box plot shows the lower quartile, and the median and upper quartile values, and the whiskers show the range of the phenotypic variation in the population.

We also detected multiple SNP associations with variation in cell wall crystallinity and fiber morphology. It has been widely acknowledged that cell wall crystallinity impacts enzymatic saccharification of the polysaccharides (Mansfield *et al.*, 1999; Chang & Holtzapple, 2000; Mosier *et al.*, 2005; R. Kumar *et al.*, 2009). Our study identified SNPs that explained *c.* 10–12% of the total genetic variance in cell wall crystallinity, which has not, to our knowledge, been investigated in any previous association studies. We identified, among others, associations between crystallinity and allelic variants of the NAC domain transcription factor *SND2*/*ANAC073* (Fig. S17) and the fasciclin-like arabinogalactan protein gene *FLA7* (Fig. S18). In *Arabidopsis*, *SND2* is a direct target of *SND1* that activates the developmental program of secondary cell wall formation. When overexpressed in *Arabidopsis*, *SND2* led to a significant increase in cell wall thickness of both the interfascicular and the xylem fibers, while RNAi-suppressed *SND2* plants showed the opposite phenotype (Zhong *et al.*, 2008). Both *SND2* and the *Eucalyptus SND2* ortholog appear to play a central role in the secondary cell wall transcriptional networks, with a key role in regulating the expression of cellulose and hemicellulose biosynthetic genes (Hussey *et al.*, 2011), but the mechanism by which allelic variation in *SND2* could affect cellulose crystallinity is not clear.

An SNP downstream of the coding region of the *FLA7* gene accounted for *c.* 4.1% of the phenotypic variance in cell wall crystallinity. *FLA7* belongs to clade A of the FLA multigene family (Fig. S18). *FLA11* and *FLA12*, which belong to a different phylogenetic subclade, are well studied in both *Arabidopsis* and *Eucalyptus* and have proposed roles in the biomechanics of the stem, which is associated with their function in cellulose deposition (MacMillan *et al.*, 2010). By contrast, the exact function of *FLA7* in *Arabidopsis* or in wood formation is currently unknown. However, a study in *Populus* showed that *FLA7* is specifically expressed in tension wood (Lafarguette *et al.*, 2004). Tension wood produces gelatinous cell wall fibers with unusual fiber properties and highly crystalline cellulose, and thus altered *FLA7* expression could affect cell wall crystallinity, suggesting potential avenues to manipulate this cellulose property.

We found significant genetic associations of multiple SNPs (including two nonsynonymous nucleotide substitutions; Fig. 4) within the *Populus FRA8* gene, which encodes a glycosyl transferase involved in xylan biosynthesis (Zhong *et al.*, 2005). A larger amount of the heritable phenotypic variance for fiber length (5.1%) was explained by the effect of the single nonsynonymous SNP scaffold_9_1380717 in *FRA8*. The *Arabidopsis fra8* ‘fragile fiber’ mutant exhibits decreased stem strength and it is known that variation in stem tensile strength is associated with variation in fiber dimensions, such as cell wall thickness and fiber length. *Populus FRA8* represents an interesting candidate for further study to examine its role in fiber development and elongation.

Solid wood properties Previous use of association genetics to dissect complex wood quality traits such as wood density and MFA in forest trees has relied on relatively small numbers of SNPs in candidate genes focusing on those directly involved in secondary cell wall biosynthesis (Dillon *et al*., 2010). Using a broader, genome-wide approach, we identified three SNPs within a *Populus* ortholog of *CYSTEINE-RICH RLK RECEPTOR-LIKE PROTEIN KINASE 42* (*CRLK42*), a cell wall-associated kinase (WAK) that individually accounted for *c*. 8% of the total genetic variance in MFA (at the most recent growth ring; Fig. S5). WAKs bind cell-wall-carbohydrate polymers (such as pectins) via extracellular binding domains and appear to modulate cell wall biosynthesis in response to environmental stimuli (Steinwand & Kieber, 2010), suggesting a possible role in modulating fiber wall properties. SNPs in the *Populus* orthologs of *STRUBBELIG-RECEPTOR FAMILY* genes *SRF3* and *SRF6* were also associated with wood density, with *SRF6* SNPs accounting for over 6% of observed phenotypic variance in this trait. The SRF family of receptor-like kinases in *Arabidopsis* are of largely unknown function, but some *SRF* genes have been hypothesized to affect different aspects of cell wall biology, and *SRF3* expression is correlated with lignification and programmed cell death in *Arabidopsis* (Eyueboglu *et al*., 2007). Finally, three SNPs in a *Populus PINORESINOL REDUCTASE 1* (*PRR1*) ortholog were significantly associated with wood density (Figs 3, S16, Table S7). *PRR1* is strongly coexpressed with secondary cell wall *CESA* genes, and an *Arabidopsis prr1* mutant shows defects in secondary cell wall biogenesis, including thinner cell walls and altered lignin content (Ruprecht *et al*., 2011). Since PRRs are implicated in the biosynthesis of lignans, phenylpropanoid compounds that may be incorporated into lignin and secondary cell walls (Nakatsubo *et al*., 2008), these genes provide intriguing new candidates for validation and further study as potentially important components of the genetic architecture of these complex traits.

Hemicellulose content and composition Hemicelluloses are believed to interact with cellulose microfibrils (Gorshkova *et al*., 2012) and these interactions can impede the access of hydrolytic enzyme to the cellulose chains. However, these noncellulosic polysaccharides are important regulators of the assembly of secondary cell walls and can impact lignification and the orientation of cellulose microfibrils (Donaldson & Knox, 2012). *Populus* wood hemicelluloses are primarily xylans and, to a lesser extent, mannans. In this study, we identified SNPs in two transcription regulators associated with hemicellulose content variation that have not been previously reported to be involved in the regulation of secondary wall formation (an AP2/B3-like transcriptional factor family protein and an SUV2 homolog with putative histone lysine methyltransferase (HMTase) activity; Table S3). We also found that variation in xylose content was associated with a gene encoding a calmodulin-binding protein IQD21 that is specifically expressed in developing *Populus* secondary xylem (Ko *et al*., 2012). An association was also identified between mannose content and an SNP in a *Populus* homolog of the *Arabidopsis FLA11* gene encoding a fasciclin-like arabinogalactan protein (Fig. S18) that is tightly coregulated with secondary cell wall *CESA* genes in *Arabidopsis* (Brown *et al*., 2005; Persson *et al*., 2005; Ma *et al*., 2007). Together with *FLA12*, *FLA11* is thought to influence wood fiber properties by interacting with the hemicellulosic matrix (MacMillan *et al*., 2010), and *fla11* and *fla12* mutants possess higher mannan content (MacMillan *et al*., 2010). Variation in relative mannose content was also associated with an SNP in a *Populus* homolog of the *Arabidopsis* NAC domain transcription factor *VASCULAR-RELATED NAC-DOMAIN1* (*VND1*). *VND1*, a homolog of *VND7*, is a transcriptional master regulator of xylem vessel formation and is up-regulated during xylem vessel element formation in *Arabidopsis* (Kubo *et al*., 2005).

Lignin content and composition Wood with a higher total lignin content typically requires more energy and/or chemical input for pulping (Stewart *et al*., 2006). Lowering lignin content and altering its composition towards a higher syringyl-to-guaiacyl (S/G) value also minimize the requirements for feedstock pretreatment and facilitate sugar release for biofuels (Studer *et al*., 2011; Mansfield *et al*., 2012). We identified a number of associations for lignin traits with SNPs located within genes that have previously been shown to be important in secondary cell wall formation, are coexpressed with genes known to be involved in this process and/or are strongly xylem up-regulated. Examples include *Populus* homologs of *Arabidopsis SKS12* (encoding a laccase), *AUD1*, *MAP20*, *WRKY32*, *ATMPK20*, *ARF-GAP domain 11* and genes encoding an FAD-binding domain-containing protein, an amino acid permease, a dirigent family protein, and a methyladenine glycosylase family protein (Tables S7,S3). Interestingly, some lignin associations were found within genes directly or indirectly related to cellulose or xylan biosynthesis. For example, MAP20 (Rajangam *et al*., 2008) associated with insoluble lignin content, while UDP-D-xylose synthase/UDP-glucuronate decarboxylase AUD1 (Persson *et al*., 2005) associated with soluble lignin content. However, given the strong genetic interrelations between the lignin biosynthetic pathway and the cellulose and hemicellulose biosynthesis pathways (Porth *et al*., 2013), this result is not surprising.

In the current study, with the exception of an association within the *CESA7-B* gene and glucose content, we did not identify any associations between trait variation in cell wall biochemistry and polymorphisms in genes dedicated to biosynthesis of cellulose or lignin (e.g. cellulose synthase or 'lignin toolbox' genes; Hamberger *et al*., 2007). This is in strong contrast to the previously conducted candidate gene-based association genetics studies in forest trees (*Eucalyptus* (Thumma *et al*., 2009), *Populus* (Wegrzyn *et al*., 2010; Guerra *et al*., 2013), pine (Gonzalez-Martinez *et al*., 2007)), where relatively small numbers of SNPs in 30–40 candidate genes were analyzed. It appears that the phenotypic variations for the wood traits that we observed across the natural range of *P. trichocarpa* are genetically controlled by polymorphisms within genes other than the commonly favored candidate genes.

Among the most important findings in our SNP association analyses is the high phenotypic variance for insoluble lignin explained by one allelic variant in *MAP20* (Fig. 5c). A recent study demonstrated that *MAP20* from hybrid aspen (*Populus tremula* × *tremuloides*) binds to cortical microtubules and is functionally linked with cellulose biosynthesis (Rajangam *et al.*, 2008). Again, these findings suggest evidence for a substantial genetic interrelation of cellulose and lignin biosynthesis (Porth *et al.*, 2013). Our study also identified a SNP within a gene encoding a FAD-binding domain-containing protein (Fig. 6) that is associated with the wood biochemistry traits insoluble lignin and total lignin content, as both traits are highly correlated (Porth *et al.*, 2013).

Conclusion

The dense marker map we applied in association mapping provided a higher chance of detecting strong LD between markers and QTLs. However, it should be stated that such an increase in the marker density does not guarantee an improvement in LD-trait detection, but conversely contributes to an increased multiple testing penalization. This is a serious limitation of association studies, especially for those investigating highly complex traits. Hence, we applied SNP preselection (Pant *et al.*, 2010), which effectively improves the power to detect small allelic effects underlying phenotypic variation (Aulchenko *et al.*, 2007). However, the verification of small effect associations across different environments and genetic backgrounds is expected to be difficult, because of the confounding effects between the genetics and the

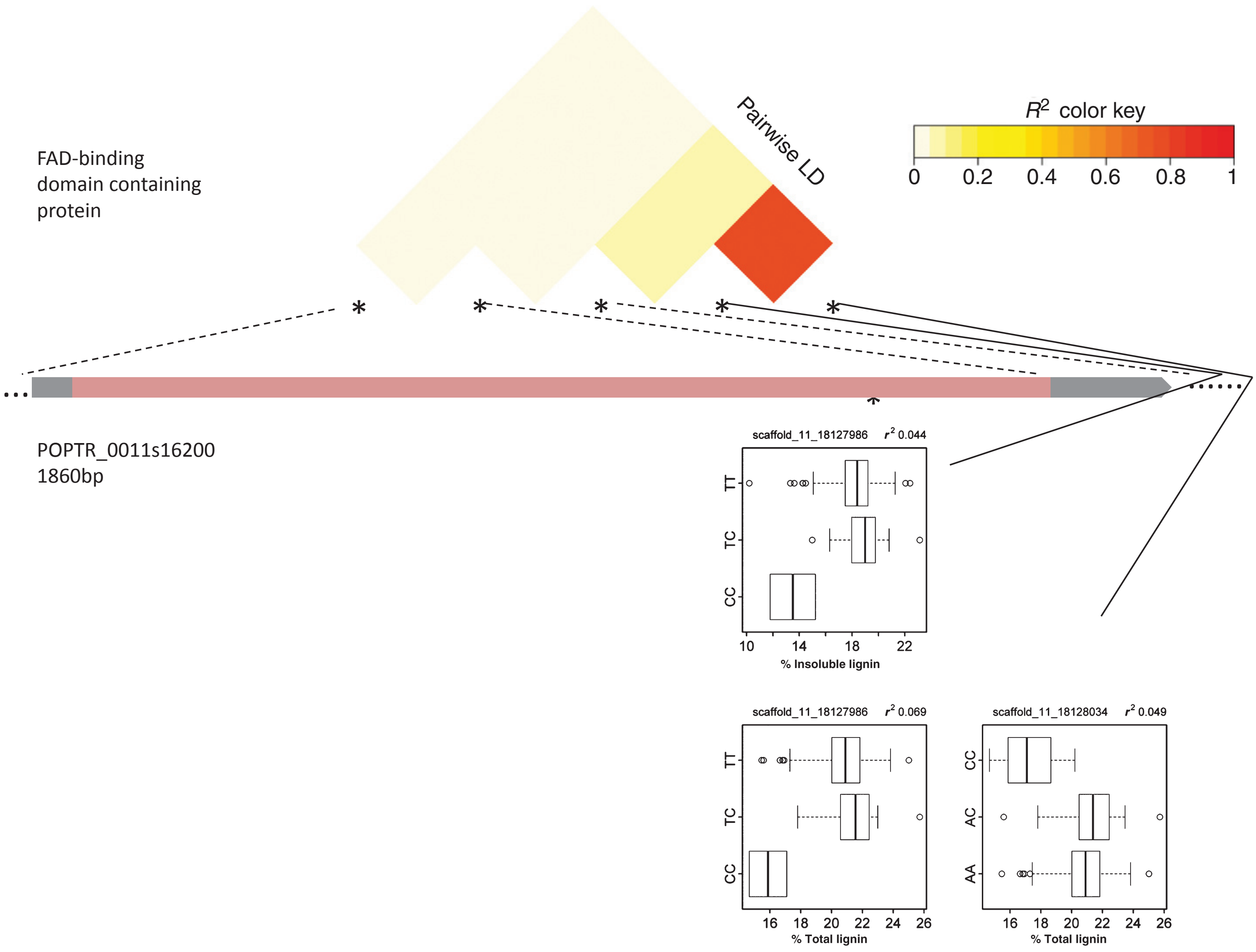


Fig. 6 Structure of the FAD-binding domain-containing protein gene (POPTR_0011s16200), linkage of genotyped single nucleotide polymorphisms (SNPs) and genetic effects of SNPs associated with insoluble lignin and total lignin content in *Populus trichocarpa*. Gene structure is represented schematically as exon (boxes)-intron (lines)-UTR (dots) patterns; the solid lines connect to the two associations ($\alpha \leq 0.05$) explaining 4.9–6.9% of relative total lignin content (percentage in extract-free dry wood) and the SNP scaffold_11_18127986, which was also significantly associated with insoluble lignin content (percentage in extract-free dry wood); dashed lines indicate other genotyped SNPs with no significant association with insoluble or total lignin content. The box in each box plot shows the lower quartile, and the median and upper quartile values, and the whiskers show the range of the phenotypic variation in the population (units in mm).

local environments, resulting in changes in gene–gene interaction networks. This phenomenon was recognized as the context-dependent effect of QTLs (Mackay *et al.*, 2009). A viable alternative to looking at one allelic effect at a time (association mapping) is the genetic tool of genomic selection (Grattapaglia & Resende, 2011), which fits all genome-wide allelic effects simultaneously and also predicts the performance of a population (pedigree or unstructured population) in terms of a particular complex trait. This allows for effective implementation in breeding activities.

Acknowledgements

This work was supported by the Genome British Columbia Applied Genomics Innovation Program (project 103BIO) and the Genome Canada Large-Scale Applied Research Project (project 168BIO), by funding to Q.C.B.C., C.J.D., R.D.G., J.E., Y.E-K., S.D.M. and by funding within the BioEnergy Science Center, a US DOE Bioenergy Research Facility supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by the University of Tennessee-Battelle LLC for the DOE.

References

Allison DB, Fernandez JR, Heo M, Zhu S, Etzel C, Beasley TM, Amos CI. 2002. Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *American Journal of Human Genetics* **70**: 575–585.

Astle W, Balding DJ. 2009. Population structure and cryptic relatedness in genetic association studies. *Statistical Science* **24**: 451–471.

Aulchenko YS, de Koning D-J, Haley C. 2007. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**: 577–585.

Brachi B, Morris G, Borevitz J. 2011. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology* **12**: 232.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635.

Brown DM, Zeef LAH, Ellis J, Goodacre R, Turner SR. 2005. Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. *Plant Cell* **17**: 2281–2295.

Carroll A, Somerville C. 2009. Cellulosic biofuels. *Annual Review of Plant Biology* **60**: 165–182.

Chang VS, Holtzapple MT. 2000. Fundamental factors affecting biomass enzymatic reactivity. *Applied Biochemistry and Biotechnology* **84–6**: 5–37.

Cronk QCB. 2005. Plant eco-devo: the potential of poplar as a model organism. *New Phytologist* **166**: 39–48.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. *Plos Biology* **8**: e1000294.

Dillon SK, Nolan M, Li W, Bell C, Wu HX, Southerton SG. 2010. Allelic variation in cell wall candidate genes affecting solid wood properties in natural populations and land taces of *Pinus radiata*. *Genetics* **185**: 1477–1487.

Dinus RJ. 2000. *Genetic modification of short rotation poplar biomass feedstock for efficient conversion to ethanol: bioenergy feedstock development program.* Oak Ridge, TN, USA: Oak Ridge National Laboratory, Environmental Sciences Division.

Donaldson LA, Knox JP. 2012. Localization of cell wall polysaccharides in normal and compression wood of Radiata Pine: relationships with lignification and microfibril orientation. *Plant Physiology* **158**: 642–653.

Eckenwalder JE. 1996. *Systematics and evolution of* Populus. National Research Council of Canada. Ottawa, ON, Canada: NRC Research Press.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11**: 446–450.

Eyueboglu B, Pfister K, Haberer G, Chevalier D, Fuchs A, Mayer KFX, Schneitz K. 2007. Molecular characterisation of the STRUBBELIG-RECEPTOR FAMILY of genes encoding putative leucine-rich repeat receptor-like kinases in *Arabidopsis thaliana*. *BMC Plant Biology* **7**: 16.

Falconer DS. 1981. *Introduction to quantitative genetics.* London, UK: Longman.

Gao X, Stamier J, Martin ER. 2008. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology* **32**: 361–369.

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**(Suppl 1): 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.

Gorshkova T, Brutch N, Chabbert B, Deyholos M, Hayashi T, Lev-Yadun S, Mellerowicz EJ, Morvan C, Neutelings G, Pilate G. 2012. Plant fiber formation: state of the art, recent and expected progress, and open questions. *Critical Reviews in Plant Sciences* **31**: 201–228.

Grattapaglia D, Resende MDV. 2011. Genomic selection in forest tree breeding. *Tree Genetics and Genomes* **7**: 241–255.

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.

Guillot G, Mortier F, Estoup A. 2005. GENELAND: a computer package for landscape genetics. *Molecular Ecology Notes* **5**: 712–715.

Hamberger B, Ellis M, Friedmann M, Souza CDA, Barbazuk B, Douglas CJ. 2007. Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*: the *Populus* lignin toolbox and conservation and diversification of angiosperm gene families. *Canadian Journal of Botany-Revue Canadienne De Botanique* **85**: 1182–1201.

Hatakeyama H, Hatakeyama T. 2010. Lignin structure, properties, and applications. *Biopolymers: Lignin, Proteins, Bioactive Nanocomposites* **232**: 1–63.

Hussey SG, Mizrachi E, Spokevicius AV, Bossinger G, Berger DK, Myburg AA. 2011. SND2, a NAC transcription factor gene, regulates genes involved in secondary cell wall development in Arabidopsis fibres and increases fibre cell area in Eucalyptus. *BMC Plant Biology* **11**: 173.

Ingvarsson PK. 2005. Nucleotide polymorphism and linkage disequilbrium within and among natural populations of European Aspen (*Populus tremula* L., Salicaceae). *Genetics* **169**: 945–953.

Ingvarsson PK, Garcia MV, Luquez V, Hall D, Jansson S. 2008. Nucleotide polymoirphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, Salicaceae). *Genetics* **178**: 2217–2226.

Keller SR, Levsen N, Olson MS, Tiffin P. 2012. Local adaptation in the flowering-time gene network of balsam poplar, *Populus balsamifera* L. *Molecular Biology and Evolution* **29**: 3143–3152.

Ko JH, Han KH. 2004. Arabidopsis whole-transcriptome profiling defines the features of coordinated regulations that occur during secondary growth. *Plant Molecular Biology* **55**: 433–453.

Ko JH, Kim H-T, Hwang I, Han K-H. 2012. Tissue-type-specific transcriptome analysis identifies developing xylem-specific promoters in poplar. *Plant Biotechnology Journal* **10**: 587–596.

Kremer A. 2011. Missing heritability and missing Fst of candidate genes: why does gene variation differ from trait variation in trees? *BMC Proceedings* **5** (Suppl 7): I1.

Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J, Mimura T, Fukuda H, Demura T. 2005. Transcription switches for protoxylem and metaxylem vessel formation. *Genes & Development* **19**: 1855–1860.

Kumar R, Mago G, Balan V, Wyman CE. 2009. Physical and chemical characterizations of corn stover and poplar solids resulting from leading pretreatment technologies. *Bioresource Technology* **100**: 3948–3962.

Kumar M, Thammannagowda S, Bulone V, Chiang V, Han K-H, Joshi CP, Mansfield SD, Mellerowicz E, Sundberg B, Teeri T *et al.* 2009. An update on the nomenclature for the cellulose synthase genes in *Populus. Trends in Plant Science* **14**: 248–254.

Lafarguette F, Leple JC, Dejardin A, Laurans F, Costa G, Lesage-Descauses MC, Pilate G. 2004. Poplar genes encoding fasciclin-like arabinogalactan proteins are highly expressed in tension wood. *New Phytologist* **164**: 107–121.

Loiselle BA, Sork VL, Nason J, Graham C. 1995. Spatial genetic-structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany* **82**: 1420–1425.

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

Macciotta NPP, Gaspa G, Steri R, Pieramati C, Carnier P, Dimauro C. 2009. Pre-selection of most significant SNPS for the estimation of genomic breeding values. *BMC Proceedings* **3**(Suppl 1): S14.

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

MacMillan CP, Mansfield SD, Stachurski ZH, Evans R, Southerton SG. 2010. Fasciclin-like arabinogalactan proteins: specialization for stem biomechanics and cell wall architecture in Arabidopsis and Eucalyptus. *Plant Journal* **62**: 689–703.

Mansfield SD, Kang K-Y, Chapple C. 2012. Designed for deconstruction – poplar trees altered in cell wall lignification improve the efficacy of bioethanol production. *New Phytologist* **194**: 91–101.

Mansfield SD, Mooney C, Saddler JN. 1999. Substrate and enzyme characteristics that limit cellulose hydrolysis. *Biotechnology Progress* **15**: 804–816.

Mosier N, Wyman C, Dale B, Elander R, Lee YY, Holtzapple M, Ladisch M. 2005. Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresource Technology* **96**: 673–686.

Nakatsubo T, Mizutani M, Suzuki S, Hattori T, Umezawa T. 2008. Characterization of *Arabidopsis thaliana* pinoresinol reductase, a new type of enzyme involved in lignan biosynthesis. *Journal of Biological Chemistry* **283**: 15550–15557.

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 concentrations. *New Phytologist* **182**: 878–890.

Novembre J, Stephens M. 2008. Interpreting principal component analysis of spatial population genetic variation. *Nature Genetics* **40**: 646–649.

Olson MS, Robertson AL, Takebayashi N, Silim S, Schroeder WR, Tiffin P. 2010. Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*). *New Phytologist* **186**: 526–536.

Pant SD, Schenkel FS, Verschoor CP, You Q, Kelton DF, Moore SS, Karrow NA. 2010. A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in holstein cattle. *Genomics* **95**: 176–182.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *Plos Genetics* **2**: 2074–2093.

Peloso GM, Lunetta KL. 2011. Chioce of population structure informative principal componentsfor adjustment in case–control study. *BMC Genetics* **12**: 64.

Peres-Neto PR, Legendre P, Dray S, Borcard D. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* **87**: 2614–2625.

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.

Pinheiro J, Bates D. 2000. *Mixed-effects models in S and S-PLUS.* New York, USA: Springer.

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

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**: 904–909.

Price AL, Zaitlen NA, Reich D, Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**: 459–463.

Rafalski JA. 2010. Association genetics in crop improvement. *Current Opinion in Plant Biology* **13**: 174–180.

Rajangam AS, Kumar M, Aspeborg H, Guerriero G, Arvestad L, Pansri P, Brown CJL, Hober S, Blomqvist K, Divne C *et al.* 2008. MAP20, a microtubule-associated protein in the secondary cell walls of hybrid aspen, is a target of the cellulose synthesis inhibitor 2,6-dichlorobenzonitrile. *Plant Physiology* **148**: 1283–1294.

Ralph SG, Chun HJE, Cooper D, Kirkpatrick R, Kolosova N, Gunter L, Tuskan GA, Douglas CJ, Holt RA, Jones SJM *et al.* 2008. Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. *BMC Genomics* **9**: 57.

Ruprecht C, Mutwil M, Saxe F, Eder M, Nikoloski Z, Persson S. 2011. Large-scale co-expression approach to dissect secondary cell wall formation across plant species. *Frontiers in Plant Science* **2**: 23.

Sannigrahi P, Ragauskas AJ, Tuskan GA. 2010. Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels Bioproducts & Biorefining-Biofpr* **4**: 209–226.

Setakis E, Stirnadel H, Balding D. 2006. Logistic regression protects against population structure in genetic association studies. *Genome Research* **16**: 290–296.

Shen H, He X, Poovaiah CR, Wuddineh WA, Ma J, Mann DGJ, Wang H, Jackson L, Tang Y, Stewart CN Jr *et al.* 2012. Functional characterization of the switchgrass (*Panicum virgatum*) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. *New Phytologist* **193**: 121–136.

Shin JH, Blay S, McNeney B, Graham J. 2006. LDheatmap: an R runction for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software* **16**: Code Snippet 3.

Shriner D. 2011. Investigating population stratification and admixture using eigenanalysis of dense genotypes. *Heredity* **107**: 413–420.

Sillanpaa MJ. 2011. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* **106**: 511–519.

Slatkin M, Muirhead CA. 1999. Overdominant alleles in a population of variable size. *Genetics* **152**: 775–781.

Slavov GT, DiFazio SP, Martin J, Schackwitz W, Muchero W, Rodgers-Melnick E, Lipphardt MF, Pennacchio CP, Hellsten U, Pennacchio LA *et al.* 2012. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa. New Phytologist* **196**: 713–725.

Stanton B, Neale D, Li S. 2010. *Populus* breeding: from the classical to the genomic approach. In: Jansson S, Rishikesh B, Groover A, eds. *Genetics and genomics of* Populus. New York, USA: Springer, 309–348.

Steinwand BJ, Kieber JJ. 2010. The role of receptor-like kinases in regulating cell wall function. *Plant Physiology* **153**: 479–484.

Stewart JJ, Kadla JF, Mansfield SD. 2006. The influence of lignin chemistry and ultrastructure on the pulping efficiency of clonal aspen (*Populus tremuloides* Michx.). *Holzforschung* **60**: 111–122.

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.

Taylor NG, Howells RM, Huttly AK, Vickers K, Turner SR. 2003. Interactions among three distinct CesA proteins essential for cellulose synthesis. *Proceedings of the National Academy of Sciences, USA* **100**: 1450–1455.

Team RDC. 2011. *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Teyssèdre S, Dupuis MC, Guérin G, Schibler L, Denoix JM, Elsen JM, Ricard A. 2012. Genome-wide association studies for osteochondrosis in French Trotter horses. *Journal of Animal Science* **90**: 45–53.

Thomas A, Abel HJ, Di Y, Faye LL, Jin J, Liu J, Wu Z, Paterson AD. 2011. Effect of linkage disequilibrium on the identification of functional variants. *Genetic Epidemiology* **35**(Suppl 1): S115–S119.

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.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

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.

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-Revue Canadienne De Recherche Forestiere* **39**: 519–526.

Xu S. 2003. The theoretical basis of the Beavis effect. *Genetics* **165**: 2259–2268.

Yuan JS, Tiller KH, Al-Ahmad H, Stewart NR, Stewart CN Jr. 2008. Plants to power: bioenergy to fuel the future. *Trends in Plant Science* **13**: 421–429.

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.

Zhong RQ, Pena MJ, Zhou GK, Nairn CJ, Wood-Jones A, Richardson EA, Morrison WH, Darvill AG, York WS, Ye ZH. 2005. Arabidopsis fragile fiber8, which encodes a putative glucuronyltransferase, is essential for normal secondary wall synthesis. *Plant Cell* **17**: 3390–3408.

Zhu C, Yu J. 2009. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* **182**: 875–888.

Supporting Information

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

Fig. S1 Geographical distribution of original sampling locations (drainages) along with levels of population structure (Geneland) in *P. trichocarpa.*

Fig. S2 Details of association data for each trait in *P. trichocarpa.*

Fig. S3 Extent of LD between the top 20 SNPs in the simple model for galactose in *P. trichocarpa.*

Fig. S4 Theoretical power of association analysis as a function of LD (r^2) between an SNP marker and a QTL explaining 0.5, 2, or 5% of the total phenotypic variance in a population of variable sizes.

Fig. S5 *CRLK42* gene structure, linkage of genotyped SNPs and genetic effects of three SNPs associated with MFA1 in *P. trichocarpa.*

Fig. S6 Boxplot diagrams depicting the genetic effects of all SNPs located within transcription factors with significant associations with relative mannose and alpha cellulose content (percentage variation in extract-free dry wood) in *P. trichocarpa.*

Fig. S7 Boxplot diagrams depicting the genetic effects of SNPs significantly associated with relative holocellulose, alpha cellulose, and hemicellulose content in *P. trichocarpa.*

Fig. S8 *ARA-5* (RabGTPase) gene structure, linkage of genotyped SNPs and genetic effects of two SNPs associated with relative holocellulose content in *P. trichocarpa.*

Fig. S9 Structure of the epsin N-terminal homology (ENTH) domain-containing protein gene (POPTR_0018s13970), linkage of genotyped SNPs and genetic effects of four SNPs associated with relative alpha cellulose content in *P. trichocarpa.*

Fig. S10 Boxplot diagrams depicting the genetic effects of SNPs significantly associated with relative glucose, xylose, mannose, and arabinose content in *P. trichocarpa.*

Fig. S11 Boxplot diagrams depicting the genetic effects of SNPs significantly associated with relative insoluble lignin, soluble lignin, total lignin, and syringyl lignin content in *P. trichocarpa.*

Fig. S12 *ATMPK20* gene structure, linkage of genotyped SNPs, and genetic effects of three SNPs associated with relative insoluble lignin content in *P. trichocarpa.*

Fig. S13 Boxplot diagram depicting the genetic effect of an SNP detected in *AUD1* with a significant association with soluble lignin content in *P. trichocarpa.*

Fig. S14 Boxplot diagrams depicting the genetic effects of SNPs significantly associated with average wood density, % cell wall crystallinity, fiber length, MFA1, and MFA2 in *P. trichocarpa.*

Fig. S15 Manhattan plot showing genome-wide SNP–trait associations for fiber length in *P. trichocarpa.*

Fig. S16 *PRR1* gene structure, linkage of genotyped SNPs, and genetic effects of three SNPs associated with average wood density in *P. trichocarpa.*

Fig. S17 *ANAC073* gene structure, linkage of genotyped SNPs and genetic effects of two SNPs associated with cell wall crystallinity in *P. trichocarpa.*

Fig. S18 Unrooted maximum likelihood phylogenetic tree of the FLA gene family.

Table S1 Information about the accessions and provenances and the phenotyping of the *Populus trichocarpa* association mapping population used in the study

Table S2 Bayesian information criterion (BIC) table for comparing the simple model (i.e. no correction for population or familial structure), the *K* model (i.e. correction for familial structure only), the PCA-TOP10 model (i.e. a PCA-based model that uses the first 10 PCs) and the PCA-BIC model (i.e. a PCA-based model that selects PCs via BIC)

Table S3 Comprehensive summary of 141 SNP associations within 105 individual genes with trait associations at $\alpha \leq 0.05$ in *P. trichocarpa*, testing 17 wood chemistry and ultrastructure traits and *c.* 3500 broad-based candidate genes

Table S4 SNP–trait associations with more than one SNP per *P. trichocarpa* gene

Table S5 SNP–trait associations in genes encoding transcription factors for *P. trichocarpa*

Table S6 SNP–trait associations in genes involved in cell wall metabolism for *P. trichocarpa*

Table S7 SNP–trait associations in genes included based on expression for *P. trichocarpa*

Table S8 Sequences of the *FLA* gene family used to build its phylogeny as displayed in Fig. S18

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.