Allelic Variation in a Cellulose Synthase Gene (PtoCesA4) Associated with Growth and Wood Properties in Populus tomentosa

Qingzhang Du,*,§,1 Baohua Xu,*,§,1 Wei Pan,*,§ Chenrui Gong,*,§ Qingshi Wang,*,§ Jiaxing Tian,*,§ Bailian Li*,§,† and Deqiang Zhang*,§,2

*National Engineering Laboratory for Tree Breeding, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, P. R. China
§Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, College of Biological Sciences and Technology, Beijing Forestry University, Beijing 100083, P. R. China
†Department of Forestry, North Carolina State University, Raleigh, NC 27695-8203 USA

Sequence data from this article has been deposited in the GenBank Data Library under the accession Nos.KC762249-KC762292

© The Author(s) 2013. Published by the Genetics Society of America.
Running Title: Allelic variation in PtoCesA4 associated with growth and wood properties

These authors contributed equally to this work

Corresponding Author: Deqiang Zhang,
National Engineering Laboratory for Tree Breeding, College of Biological Sciences and Technology, Beijing Forestry University, No. 35, Qinghua East Road, Beijing 100083, P. R. China.
Tel: +86-10-62336007; Fax: +86-10-62336164
Email: DeqiangZhang@bjfu.edu.cn

Key words: Linkage disequilibrium (LD)-linkage analysis, Multi-locus association models, Populus tomentosa, RNA transcript analysis, Single nucleotide polymorphism.
ABSTRACT

Lignocellulosic biomass from trees provides a renewable feedstock for biofuels, lumber, pulp, paper, and other uses. Dissecting the mechanism underlying natural variation of the complex traits controlling growth and lignocellulose biosynthesis in trees can enable marker-assisted breeding to improve wood quality and yield. Here, we combined linkage disequilibrium (LD)-based association analysis with traditional linkage analysis to detect genetic effect of a *Populus tomentosa* cellulose synthase gene, *PtoCesA4*. *PtoCesA4* is strongly expressed in developing xylem and leaves. Nucleotide diversity and LD in *PtoCesA4*, sampled from the *P. tomentosa* natural distribution, revealed that *PtoCesA4* harbors high single nucleotide polymorphism (SNP) diversity ($\pi_T = 0.0080$ and $\theta_w = 0.0098$) and low LD ($r^2 \geq 0.1$, within 1400 bp), demonstrating that the potential of a candidate-gene-based LD approach in understanding the molecular basis underlying quantitative variation in this species. By combining single-SNP, multi-SNP and haplotype-based associations in an association population of 460 individuals, with single-SNP linkage analysis in a family-based linkage populations (1200 individuals), we identified three strong associations (false discovery rate $Q < 0.05$) in both populations. These include two non-synonymous markers (SNP49 associated with $\alpha$-cellulose content and SNP59 associated with fiber width) and a noncoding marker (SNP18 associated with $\alpha$-cellulose content). Variation in RNA transcript abundance among genotypic classes of SNP49 was confirmed in these two populations. Therefore, combining different methods allowed us to examine functional *PtoCesA4* allelic variation underlying natural variation in complex quantitative traits related to growth and lignocellulosic biosynthesis.
INTRODUCTION

Wood formation represents a major carbon sink for the biosphere, as well as providing an important renewable resource for lumber, pulp and paper, and possibly for biofuels feedstocks (Li et al. 2006). Wood formation mainly includes deposition of strong secondary cell walls that contain cellulose microfibrils, lignin, and other components. Many studies have examined the molecular biology of secondary cell wall biosynthesis, and have shown that the complex, dynamic process of secondary wall formation requires the coordinate regulation of the diverse metabolic pathways involving polysaccharides and lignin (Persson et al. 2005; Somerville 2006). As a biomaterial, wood varies in its properties; specific compositions and structural features make wood more suited for different applications. For example, high-lignin wood can release more thermal energy and thus would be well-suited for thermochemical biofuels applications. Moreover, wood yield is another important trait for trees used as crop species for lumber or biofuels feedstocks. Variation in wood properties likely depends on variation in genes involved in xylogenesis (Neale and Savolainen 2004; Li et al. 2009), making these traits amenable to a candidate-gene approach. Few functional studies in forest trees have identified genes directly affecting wood quality (Spokevicius et al. 2007), largely due to their long generation intervals, large size, and lack of mutant libraries for reverse genetics (González-Martínez et al. 2006; Zhang et al. 2010a).

Association studies are an effective means to bridge the gap in our understanding between complex, quantitative traits and the underlying genetic variation at specific
candidate genes or multiple loci dispersed genome-wide (Sexton et al. 2012). A diverse
group of growth and wood properties have been studied in forest tree species by a
candidate-gene approach (Thumma et al. 2005, 2009; González-Martínez et al. 2007;
Wegrzyn et al. 2010; Sexton et al. 2011; Beaulieu et al. 2011; Dillon et al. 2010, 2012;
Guerra et al. 2012). However, LD mapping may generate false positives due to population
structure (Atwell et al. 2010), although statistical methods to control for population
structure have been developed (Yu et al. 2006; Shriner et al. 2007). Results from
association studies in these species must be cautiously evaluated and, ideally, verified or
supported by other approaches, such as quantitative trait locus (QTL) linkage analysis,
transgenesis, or transcriptome profiling (Manenti et al. 2009; Ingvarsson and Street 2011).

Conventional QTL linkage analysis in controlled crosses and alternative LD-based
association mapping using diverse germplasms are two broadly used approaches for the
dissection of the genetic architecture of complex traits (Brachi et al. 2010; Sterken et al.
2012). A linkage approach is powerful for detecting genetic effects at loci involved in the
expression of target traits, which often identifies large chromosome regions of interest with
relatively low marker coverage, due to QTL mapping uses only the recombination
information found in the progeny of two parents. By contrast, LD mapping offers the
ability to exploit all recombination events that have occurred in the evolutionary history of
a sample set of germplasm, allowing for increased mapping resolution with either prior
information on candidate genes or a genome-wide scans with very high marker coverage
(Manenti et al. 2009; Lu et al. 2010). The complementary use of traditional linkage
mapping and LD-based association mapping would further improve mapping resolution
without requiring dense marker maps, by combining the advantages, and overcoming some of the inherent limitations, of both approaches (Myles et al. 2009; Lu et al. 2010). This integrated strategy enables a closer examination of the number and effect sizes of genes responsible for traits of interest through complex trait dissection in several plant species (Thumma et al. 2005, 2009; Stich and Melchinger 2009; Lu et al. 2010; Brachi et al. 2010).

Here we examined the number and effect magnitudes of allelic polymorphisms in a candidate gene underlying natural variation of growth and wood properties, using integrated linkage-LD mapping.

Cellulose is the major component of secondary cell walls; its biosynthesis is catalyzed by cellulose synthases (CesA), located in the plasma membrane (Suzuki et al. 2006). The catalytic subunits of the cellulose synthesizing complexes are encoded by the CesA gene family, and different sets of CesAs dominate cellulose synthesis in primary and secondary cell walls (Taylor et al. 2003; Burton et al. 2004; Persson et al. 2007). For example, Arabidopsis CesAs (AtCesA4, AtCesA7, and AtCesA8) are involved in cellulose biosynthesis of the secondary walls (Atanassov et al. 2009). Poplars (Populus spp.) have a long tradition as a model system for studies of angiosperm tree physiology and genetics. CesA homologs have been identified in Populus and they also have been used to investigate the mechanisms of cellulose biosynthesis (Suzuki et al. 2006; Kumar et al. 2009). The first tree CesA gene was isolated from aspen (P. tremuloides) by Wu et al. (2000). Since then, 17 CesA family members have been identified in aspen and its hybrids (P. tremula × tremuloides) (Djerbi et al. 2004). 18 CesAs (encoding 17 proteins) have been identified in P. trichocarpa (Suzuki et al. 2006), and five P. trichocarpa CesAs (PtiCesA4, PtiCesA7-A
and -B, and *PtiCesA8*-A and -B) are expressed in developing xylem tissue undergoing secondary wall thickening (Kumar *et al.* 2009).

Here, candidate gene approaches were used to examine genetic variation in only one of the *Populus tomentosa* cellulose synthase gene homologs, *PtoCesA4*, underlying complex quantitative traits related to growth and lignocellulosic biosynthesis. On the one hand, we use a combination of single-SNP models, multi-SNP models, and haplotype-based association methods in an association population (460 individuals), with single-SNP linkage analysis in a family-based linkage population (1200 individuals) to identify several associations underlying natural variation of important wood properties. On the other hand, to probe the mechanism of this variation, we also examine *PtoCesA4* expression, finding that it is expressed in developing xylem and that its expression varies in lines with different genotypes for one of the associated markers.

**MATERIALS AND METHODS**

Population materials and phenotypic data

**Association population:** An association population of 460 unrelated *P. tomentosa* individuals representing all of the original provenances in the entire natural distribution region of *P. tomentosa* (30–40°N, 105–125°E), were used for initial SNP association (Du *et al.* 2013). The distribution zone from which these individuals were collected can be divided into three large climatic regions, southern (S), northwestern (NW), and northeastern (NE), on the basis of a principal components analysis and isodata fuzzy clustering using 16
meteorological factors (Huang 1992). 40 individuals were randomly selected from this
association population and used to identify SNPs within *PtoCesA4* using a direct
sequencing method.

**Linkage population:** The hybrid population used for linkage analysis consists of 1,200
individuals, randomly selected from 5,000 F1 progeny of controlled crosses between two
elite poplar parents (members of the section *Populus*), clone “YX01” (*P. alba* × *P.
glandulosa*) as the female and clone “LM 50” (*P. tomentosa*) as the male. These two related
species are members of the section *Populus* in the genus *Populus*. The progeny were grown
in 2008 in the Xiao Tangshan horticultural fields of Beijing Forestry University, Beijing,
China (40°2′N, 115°50′E) using a randomized complete block design with three clonal
replications (Du et al. 2012).

**Phenotypes:** All individuals of these two populations were scored for nine growth and
wood property traits, with at least three ramets per genotype. These nine traits included tree
height (H), diameter at breast height (D), stem volume (V), fiber length (FL), fiber width
(FW), microfiber angle (MFA), holocellulose, α-cellulose, and lignin content. The growth
traits, including tree height (H), diameter at breast height (D), and stem volume (V), were
measured during field surveys in 2009 using the methods described by Zhang et al. (2006).
Wood chemical compositions (holocellulose, α-cellulose, and lignin contents) were
determined using near-infrared reflectance spectroscopy (NIRS) according to Schimleck *et
al.* (2004), based on training sets (models) derived from wet chemistry analyses techniques
described in Tian *et al.* (2012). Fiber length and width were measured using the Colour
CCTV Camera (Panasonic SD II), MFA was measured by X-ray powder diffractometer
(Philips, Eindhoven, The Netherlands), and the X-ray diffraction profile was integrated at Chi between -180° and +180°. Analysis of variance (ANOVA) and phenotypic correlations for these nine traits in these two populations were reported by Tian et al. (2012) and Du et al. (2013), respectively.

Isolation of PtoCesA4 cDNA
Developing xylem tissues were collected by scraping the thin (approximately 1.0 mm) and partially lignified layer on the exposed xylem surface at the bottom stems of 1-year-old P. tomentosa clone “LM50”. These tissues were immediately frozen in liquid nitrogen, and then stored in the laboratory at -80°C for later RNA extraction. The P. tomentosa stem developing xylem cDNA library was constructed using the Superscript k System (Life Technologies, Rockville, MD), as part of our effort to identify genes expressed predominantly in the P. tomentosa stems. The details of constructing the cDNA library were previously described by Li et al (2009). The constructed cDNA library consisted of 5.0×10^6 pfu with an insert size range of 1.0–4.0 kb. Random end-sequencing of 1,000 cDNA clones and comparison with Arabidopsis or P. trichocarpa CesA sequences identified a full-length cDNA with high similarity to AtCesA4 (73.6%) or PtiCesA4 (97.4%). Therefore, we named this cDNA PtoCesA4.

DNA extraction, PtoCesA4 identification, and phylogenetic analysis
Total genomic DNA was extracted from young leaves with the DNeasy Plant Mini kit (Qiagen China, Shanghai). Specific primers were designed for sequencing PtoCesA4 based
on cDNA sequence. 6,421 bp of genomic DNA sequences for \textit{PtoCesA4}, including the promoter (1,111 bp), were obtained by direct sequencing in the \textit{P. tomentosa} LM50 clone, using conserved primers, and the BigDye Terminator Cycle Sequencing kit, version 3.1 (Applied Biosystems, Beijing, China), run on a Li-Cor 4300 genetic analyzer (Li-Cor Biosciences, Lincoln, Nebraska, USA). The \textit{PtoCesA4} sequence was deposited in GenBank under the accession number KC762249.

To analyze the phylogenetic relationship of \textit{PtoCesA4} to the CesAs from other species, the amino acid sequences of CesA from \textit{Arabidopsis thaliana}, rice (\textit{Oryza sativa}), maize (\textit{Zea mays}), \textit{Eucalyptus grandis}, and \textit{Populus trichocarpa} were identified from NCBI (http://www.ncbi.nlm.nih.gov) using BLAST (Altschul et al. 1997). Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 4, and the neighbor-joining (NJ) method was used to build phylogenetic trees (Tamura et al. 2007). Statistical confidences of the nodes of the tree are based on 1000 bootstrap replicates.

\textit{Tissue-specific expression analysis of PtoCesA4}

\textbf{RNA extraction and cDNA synthesis:} Total RNA was extracted from various tissues, including root, stem (bark, phloem, cambium, developing xylem, mature xylem), mature leaf, and apical shoot meristem of 2-year-old \textit{P. tomentosa} clone “LM50”, with the Qiagen RNAeasy kit (Qiagen China, Shanghai). Additional on-column DNase digestions were performed three times during RNA purification using RNase-Free DNase (Qiagen). RNA was quantified based on absorption at 260 nm. Quantified RNA was reverse-transcribed into cDNA with the SuperScript First-Strand Synthesis system and the supplied
polythymidylate primers (Invitrogen). All cDNA samples were used for tissue-specific expression analysis of PtoCesA4 using Real-time quantitative PCR (RT–qPCR).

**Real-time quantitative PCR:** Quantitative PCR was carried out on a DNA Engine Opticon 2 machine (MJ Research) using the LightCycler-FastStar DNA master SYBR Green I kit (Roche). The Real-time quantitative PCR and the generated real-time data analysis were performed following Zhang *et al.* (2010a). The PtoCesA4-specific (F: 5’-GCCAGTCTGCAACGTCGAA-3’; R: 5’-GGAAAGCCACACACA TGAC-3’) and internal control (Actin) primer pairs (F: 5’-CTCCATCATGAAATGCGATG-3’; R: 5’-TTGGGGCTAGTGCTGAGATT-3’) were designed using Primer Express 3.0 (Applied Biosystems). All reactions were performed in triplicate technical and triplicate biological repetitions, respectively. The results obtained for the different tissues were standardized to the levels of Actin.

**Transcript analysis of SNP genotypes**

Transcript levels were determined for SNP genotypes significantly associated with phenotypic traits, to test whether transcript abundance varied in the different SNP genotypic classes. Only significant SNPs (false discovery rate FDR ($Q \leq 0.10$) were targeted in the association and linkage populations, respectively (see Results). Transcript levels were determined by RT–qPCR with gene-specific primers. For each SNP genotypic class, 10 trees were individually sampled by obtaining secondary xylem at 1.3 m from the ground; tissue handling and RNA extractions were performed as described above. The differential expression across three or two genotypic classes was tested by ANOVA.
SNP discovery and genotyping

The *PtoCesA4* gene, including 1.111 kb of the promoter, was sequenced and analyzed in 40 unrelated individuals from the *P. tomentosa* association population to identify SNPs, without considering insertions/deletions (INDELs). Sequencher v.4.0 and BioEdit were used for sequence alignment, and manual editing was used to confirm sequence quality and to remove primer sequences. To identify putative SNP variants, eight clones for each individual were randomly picked for initial allele sequencing on a Li-Cor 4300 genetic analyzer (Li-Cor Biosciences, Lincoln, Nebraska, USA). Alignments and SNP discovery for this gene among 40 unrelated individuals described here are based on the *P. trichocarpa* genome ([http://genome.jgi-psf.org/Poptr1/Poptr1.home.html](http://genome.jgi-psf.org/Poptr1/Poptr1.home.html), Tuskan et al., 2006). All 40 sequences have been deposited in GenBank (accession nos. KC762252–KC762291). Subsequently, 92 common SNPs (minor allele frequencies > 0.10; Supporting Information, Table S1) were genotyped by single-nucleotide primer extension using a Beckman Coulter sequencing system for all DNA samples.

Nucleotide diversity and linkage disequilibrium

Nucleotide diversity: Diploid sequences were disambiguated into haplotypes using Phase v. 2.1 using 10,000 iterations of the Bayesian MCMC chain (Stephens and Scheet 2005). We then used the phased haplotypes to estimate the number of segregating sites, nucleotide diversity, and neutrality test. The DnaSP program version 4.90.1 (Rozas *et al*. 2003) was used to calculate summary statistics for SNP polymorphisms. Nucleotide diversity was
estimated using both the average number of pair-wise differences per site between
sequences, \( \pi \) (Nei 1987), and the average number of segregating sites, \( \theta_w \) (Watterson 1975).
Diversity statistics were also calculated separately for noncoding, synonymous and
nonsynonymous sites. Neutrality test statistics, Tajima's D (Tajima 1989) and Fu and Li's D
(Fu and Li 1993), were calculated separately for three climatic regions and the complete
data set, and tested using 10,000 simulations to test whether a gene or genomic region is
evolving randomly (neutral evolution) or whether the region is under selection (non-neutral
evolution), and the statistical significance of Tajima’s D was determined using the software
DnaSP version 4.90.1.

**Linkage disequilibrium**: Linkage disequilibrium (LD) was measured as the squared
correlation of allele frequencies \( r^2 \), which is affected by both recombination and differences
in allele frequencies between sites (Hill and Robertson 1968). The \( r^2 \) value between each
pair of common SNPs (minor allele frequencies > 0.10) in the candidate gene was
calculated with \( 10^5 \) permutations, using TASSEL Ver. 2.0.1 ([http://www.maizegenetics.net/](http://www.maizegenetics.net/)).
To assess the extent of LD within the sequenced region of *PtoCesA4*, the decay of LD
within a specific physical distance (base pairs) between common SNP sites within this gene
was estimated by nonlinear regression (Remington *et al.* 2001). This analysis was done
both within three climatic regions and for the complete data set. Singletons were excluded
in the LD analyses.

**Marker-trait association analysis**

**Single-SNP models**: The unified mixed linear model (MLM) was used for single SNP-trait
analysis, with $10^4$ permutations in TASSEL Ver. 2.0.1 (Yu et al. 2006; Bradbury et al. 2007). These phenotypes were centered and standardized before analysis. In this MLM (Q + K model) described previously, the population structure matrix ($Q$) was identified based on the significant subpopulation structure in this association population ($K = 11$; Du et al., 2012). The relative kinship matrix ($K$) has been obtained using the method proposed by Ritland et al. (1996) in Du et al. (2013). Corrections for multiple testing were performed using the positive false discovery rate (FDR) with $10^4$ permutations in QVALUE (Storey and Tibshirani 2003). The modes of gene action were quantified using the ratio of dominance ($d$) to additive ($a$) effects estimated from least-square means for each single SNP association. Details of the algorithm and formulas for calculating gene action were previously described (Eckert et al. 2009; Wegrzyn et al. 2010).

**Multi-SNP models:** Bayesian linear mixed models incorporating effects of population structure were used to construct multi-locus models for each trait (Quesada et al. 2010; Eckert et al. 2011). These phenotypes were centered and standardized before analysis. For each trait, multilocus models were subsequently constructed from the list of SNPs with significant effects ($P < 0.05$). Model parameters, including 95% credible intervals for SNP effects, were estimated using Markov chain Monte Carlo (MCMC) with 50,000 steps after an initial burn-in of 10,000 steps. All linear mixed-model analyses were conducted using the Bayesian association with the missing data (BAMD) program in R (http://cran.r-project.org/package=BAMD).

**Haplotype analysis:** On the basis of the information from the LD blocks surrounding the significant SNPs ($P < 0.05$, Table S2), the haplotype (a block of linked ordered markers)
frequencies of loci genotypes were estimated based on genotypic data of 460 individuals, and haplotype-based association tests with growth and wood-quality traits were performed, using FAMHAP ver. 19 (http://famhap.meb.uni-bonn.de/index.html). FAMHAP estimates haplotype frequencies using maximum-likelihood. Singleton alleles were ignored when constructing the haplotypes, and haplotypes with a frequency <5% were also discarded. The input consisted of genotype matrices with structure analysis matrices (Q) and phenotypic value matrices, and significances of the haplotype associations were identified based on $10^4$ permutation tests. A correction for multiple testing was performed using the positive FDR.

**Single-SNP linkage analysis:** Comparing the $PtoCesA4$ sequences in parents of this linkage population (accession nos. KC762249–KC762252), we identified a panel of SNP markers, which was based on the common SNPs detected in the association population (Table S1). Inheritance tests of all SNPs were first examined in the linkage population with 1200 individuals, by performing a chi-square ($\chi^2$) test at 0.01 probability, and then SNPs following Mendelian expectations ($P \geq 0.01$), were used in single-marker analysis in the linkage mapping population (excluding the genotype data involving null alleles at each locus). Significant SNPs were calculated by fitting the data to the model $y = \mu + m_i + e_{ij}$, where $y$ is the trait value, $\mu$ is the mean, $m_i$ is the genotype of the $i$th marker, and $e_{ij}$ is the residual associated with the $j$th individual in the $i$th genotypic class. Percent phenotypic variance explained by the most significant marker was calculated, and the FDR method was used to perform a correction for multiple testing.
RESULTS

Identification and phylogenetic analysis of PtoCesA4

We used reverse transcription (RT)-PCR to isolate a full-length cDNA of PtoCesA4 from a cDNA library prepared from the developing xylem zone of P. tomentosa. The cDNA clone PtoCesA4 (GenBank Accession No. KC762292) is 3,757 bp in length, and the open reading frame (3,129 bp) encodes a polypeptide of 1,042 amino acids with an estimated molecular mass of 118.4 kD and a pI of 7.60, flanked by 297 bp of 5' untranslated leader region (5'UTR) and 331 bp of 3' UTR (Figure 1). Nucleotide sequence comparison of PtoCesA4 cDNAs with known full-length Arabidopsis CesA cDNA sequences, revealed that PtoCesA4 is a member of the CesA gene family because it contains all of the conserved features (Holland et al. 2000; Chen et al. 2010), such as a putative zinc-binding domain (at amino acid residues 31-76), two transmembrane helices (at residues 217-238 and 250-267) in the N-terminal region and six in the C-terminal region (Figure 1).

To investigate the evolutionary relationship between the PtoCesA4 and other eukaryotic CesAs, including genes from monocots (rice and maize), Arabidopsis, Eucalyptus grandis, and black cottonwood, an unrooted tree was generated from 37 full CesA protein sequences using 1,000 replication bootstrap values (Figure S1). Phylogenetic analysis revealed that the cloned PtoCesA4 is an ortholog of the AtCesA4 and PtiCesA4 (Figure S1).

Tissue-specific expression patterns of PtoCesA4

We first determined to what extent PtoCesA4 exhibited xylem-specific expression. Levels
of *PtoCesA4* mRNA in various poplar tissues, including root, bark, phloem, cambium, developing xylem, mature xylem, mature leaf, and apical shoot meristem, was examined by RT–qPCR with gene-specific primers and *Actin* as an internal control. *PtoCesA4* transcripts were present in all plant organs, including root, stem, and leaf, with varying patterns of expression (Figure 2 and File S1). In leaf and root, *PtoCesA4* is most abundant in mature leaf (0.8513). Low abundance is observed in apical shoot meristem and root (0.1092 and 0.1247). In the stem, *PtoCesA4* shows the highest abundance in the developing xylem tissue (0.3490), followed by the mature xylem (0.3023); it has moderate abundance in the primary tissues of the bark (0.0838) and phloem (0.0387), and the lowest is found in the cambium (0.0048). Collectively, *PtoCesA4* expression in secondary tissues (xylem) was at least 60-fold higher than in primary tissues (cambium) (Figure 2), suggesting that *PtoCesA4* may be a highly expressed gene associated with secondary wall formation. The highest transcript level of this gene found in the mature leaf (Figure 2), suggesting that it may participate in shared pathways for assimilating the products of photosynthesis into sugars and starch, synthesize cell wall biopolymers, and the creation of various glycosylated compounds (Persson *et al.* 2005; Geisler-Lee *et al.* 2006).

**Nucleotide diversity in PtoCesA4**

To characterize the intraspecific molecular evolution of the *PtoCesA4*, an approximately 6421 bp genomic region of *PtoCesA4*, including 1111 bp of promoter region, 297 bp of 5'UTR, 3129 bp of exons, 1553 bp of intron, and 331 bp of 3'UTR, was amplified and sequenced from 40 unrelated individuals in a discovery population, which encompassed
most of the natural range of *P. tomentosa*. After definition of phased haplotypes among these 40 unrelated individuals using Phase v. 2.1, a more detailed analysis of SNP variation was conducted over different regions of *PtoCesA4*, and the profile of nucleotide diversity at these loci was calculated (Table 1). On the basis of the aligned sequences for 40 samples, 218 SNPs were detected in *PtoCesA4*, with a high frequency of 1/29 bp (Table 1). The highest level of nucleotide polymorphism was in the promoter region, whereas the lowest was found in the exons, as expected if the coding region is conserved under selective pressure. Fifty-five SNPs were found in exons; of these, eight led to nonsynonymous changes, and the other 47 SNPs were synonymous mutations (Table 1). In total, 210 SNPs were categorized as silent sites; 92 of the 218 SNPs (42.2%) were considered common (frequency > 0.10, Figure 1 and Table S1).

The *PtoCesA4* locus has high nucleotide diversity, where $\pi_T = 0.0080$ and $\theta_w = 0.0098$ (Table 1). The average levels of nucleotide diversity ($\pi$) were 0.0124 (silent sites), 0.0040 (exons), and 0.0065 (introns). Both $\pi$ and $\theta_w$ were higher in non-coding than in coding regions (Table 1). In coding regions, the average levels of nucleotide diversity for nonsynonymous polymorphisms ($d_N$, $\pi = 0.0009$ and $\theta_w = 0.0010$) were roughly 6-fold lower than for synonymous polymorphisms ($d_S$, $\pi = 0.0149$ and $\theta_w = 0.0194$). The $d_S/d_N$ for the exon regions was significantly less than one, reflecting the action of purifying selection at the non-synonymous sites in exons. Of the 218 single-base changes, 150 (68.8%) were transitions and 68 (31.2%) were transversions, and the ratio of transitions to transversions for these SNPs was about 2.17. Furthermore, for synonymous polymorphisms (47) in exons, 39 (80.1%) were transitions, indicating that translational selection has shaped synonymous
codon usage.

Genetic differentiation within and among three geographically independent climatic regions were studied using the nucleotide diversity data from *PtoCesA4* (Table 2). Levels of nucleotide variation (measured using $\pi$) in the three climatic regions varied but showed similar patterns of $\pi_{\text{tot}}$, $\pi_{\text{sil}}$, $\pi_s$, and $\pi_n$ (Table 2), suggesting that the level of selective constraint was similar between the climatic regions. Tajima's D was positive in the Southern and Northwestern climatic regions as opposed to negative found in the Northeastern region and the *P. tomentosa* population as a whole, but no significant departures from the neutral expectation were observed (Table 2). The Fu and Li’s D statistical tests were positive for Northwestern region, but were negative for the other regions and the *P. tomentosa* population as a whole (Table 2), revealing an excess of low-frequency mutations for this gene region in the *P. tomentosa* species-wide samples.

**Linkage disequilibrium**

Using genotypic data for 92 common SNPs located in *PtoCesA4*, the $r^2$ values were pooled to assess the overall behavior of LD within the *PtoCesA4* gene. The average value of $r^2$ is 0.45 for all SNPs within the *PtoCesA4* region, with a range from 0.0 (equilibrium) to 1.0 (disequilibrium). Figure S2 shows several high-LD distinct haplotype blocks ($r^2 > 0.75$, $P < 0.001$) across the sequenced regions. Especially, a higher LD level among physically linked loci (the longer haplotype block) is present in the promoter region, compared with the other regions (Figure S2). There are a larger number of markers that are in linkage equilibrium among these distinct haplotype blocks ($r^2 < 0.3$; Figure S2). The nonlinear regression
shows a clear and rapid decline of LD with distance in base pairs within *PtoCesA4* (*r^2* ≥ 0.1, within 1400 bp, Figure 3), indicating that LD did not extend over the entire gene region. Nevertheless, within-group analyses of LD show a slightly higher level of LD within each geographical climatic region with the *r^2* values declining to 0.1 within c. 2800 bp (Southern and Northwestern regions) and c. 3200bp (Northeastern region) (Figure 3).

Detection of phenotype-genotype associations

**Single SNP-based associations:** A total of 828 tests (92 SNPs × 9 traits, File S2 and S3) were conducted with 10^4 permutations using MLM. In all, 41 significant associations with all nine phenotypic traits were identified at the threshold of *P* < 0.05, representing 24 SNPs from different regions within *PtoCesA4* (Table S2). Corrections for multiple testing using the FDR method reduced these 41 associations to 14 (*Q* < 0.10, Table 3). These 14 associations representing ten unique SNPs from the promoter, exon, intron, and 5′UTR regions of *PtoCesA4*, were significantly associated with seven phenotypic traits, excluding fiber length and microfibril angle traits (Table 3). These loci explained a small proportion of the phenotypic variance, ranging from 1.6% to 5.3% (Table 3), in accordance with polygenic quantitative models of wood traits (Neale and Savolainen 2004; Beaulieu *et al.* 2011).

Of these ten unique SNPs, there were 2 nonsynonymous, 2 synonymous, and 6 noncoding SNPs (Table 3). The non-synonymous marker SNP49 in exon 3, results in an amino acid change from His to Asn, associated significantly with multiple traits, i.e., α-cellulose (5.3%), lignin (3.0%), and H (2.3%). In this case, the mode of gene action
seems additive with the minor allele (A) conferring a lower lignin content and higher values in α-cellulose and H (Table 3). For the other non-synonymous marker SNP59 in exon 6, which has the minor allele (A), results in an amino acid change from Ser to Tyr, associated significantly with fiber width ($R^2 = 2.6\%$). The genotypic effects on fiber width were significant (22.38 µm in AA, 23.21 µm for AC, and 23.99 µm for CC), consistent with the additive effect of gene action on fiber width (Table 3 and Figure 4). Also, a synonymous marker SNP45 in exon 2, associated with holocellulose content, showed a difference among three genotypic classes (two significant) (74.62% in AA, 74.55% in AC, and 73.30% in CC), indicating that patterns of gene action consistent with dominant effects (Table 3). SNP75 in exon 10, the other synonymous mutation, associated with D and V, explaining 3.2% and 2.6% of the phenotypic variance, respectively (Table 3).

Of the remaining noncoding markers, SNP3 and SNP18 from the promoter region and SNP41 from the 5′UTR were significantly associated with α-cellulose content, with the small single SNP effects ranging from 1.6% to 2.5% (Table 3). SNP81 in intron 10, was significantly associated with holocellulose and D traits, and this marker has same allelic effects in these two traits: heterozygous trees (CT) for this marker showed intermediate average holocellulose content (73.78% in CT vs. 74.03% and 73.42% in CC and TT, respectively); an additive effect of gene action appeared in the D trait (20.26cm in CC, 21.43cm in CT and 22.31cm in TT). SNP44 in intron 1 and SNP48 in intron 2 were associated with lignin content and D, respectively. Four of the 10 SNP markers exhibited significant associations with at least one trait, suggesting a pleiotropic effect of these loci (Table 2).
Multi-SNP associations: Application of Bayesian linear mixed models, where each trait was evaluated against multi-SNP models, identified a multitude of new genetic associations (Table S3). In total, 38 associations were obtained across all traits representing 9 unique growth and wood properties and 22 unique SNPs. Of the 22 unique SNPs, three were non-synonymous, four were synonymous, and 15 were noncoding (Table S3). Ten of these SNPs were associated with more than one trait (range 1-4), which is likely to be due to the strong correlation between some wood and growth property traits. Effect sizes for SNPs identified with the multilocus models, when analyzed using a single locus test, were nearly two-fold lower than those detected only in the single marker models (average $R^2 = 0.015$).

All of the SNPs identified as significant in the single locus tests (FDR $Q < 0.10$) were revealed with significant effects in multi-SNP models (Tables 3 and S3). The number of SNPs retained in these models ranged from 2 (H) to 7 ($\alpha$-cellulose and D), with a mean of 4 SNPs per trait, explained larger portions of genetic effects for many traits, ranging from 3.9% to 12.4% (Table S3, Figure 5).

Haplotype-based associations: In this low LD tree species, haplotype-based association tests were performed to identify significant haplotypes with growth and wood-quality traits (Table 4). A total of 68 sets (based on the significant single SNPs at the threshold of $P < 0.05$, Table S2) were analyzed with each of the nine traits, and the number of common haplotypes (frequency > 1%) per set varied from 2 to 10 with an average of 5.0. Seventeen significant regions including 80 common haplotypes were identified at the significance threshold of $P < 0.05$ (details not shown). Multiple test corrections reduced this number to 11, at a significance threshold of $Q < 0.10$, and 21 significant haplotypes were associated
with the five phenotypic traits excluding MFA and H phenotypes (Table 4).

Most of significantly haplotype-based associations were trait-specific, but some were shared among the traits. For instance, several haplotypes from SNPs 44-46 were simultaneously associated with lignin and holocellulose, which are supported by significant single SNP associations ($Q < 0.10$, Tables 3 and 4). SNPs 89-91, in the 3'UTR region, were associated with fiber length without a supporting single marker association (Table 4), and significant differences among haplotypes were observed for this trait (1.180 mm in G-G-A and 1.501mm in C-A-A). Each haplotype explained a small proportion of phenotypic variation, from 2.6% to 5.6%, and many were strongly supported by single-SNP associations (Tables 3 and 4).

**Single-SNP linkage analysis:** Based on all 92 common SNPs detected by association mapping (Table S1), we observed 56 SNPs from the *PtoCesA4* in the linkage population, including six novel alternative SNPs identified in the parents of this hybrid population, and 50 corresponding to the positions of common SNPs detected in the association population (see Table S1). Of these SNPs, 46 markers segregated in the 1200 progeny, with a segregation ratio close to 1:2:1 for 26 SNP loci, 1:1 for 20 loci (Table S1); and the 10 significant SNP markers ($Q < 0.10$; Table 3) identified in the association population were involved in this single-SNP linkage analysis. Therefore, 414 single-marker analyses (46 SNPs × 9 traits, File S4 and S5) were conducted in this linkage mapping population. In all, eighteen associations were first observed at the threshold of $P < 0.05$ (Table S4). However, a multiple test correction reduced this number to seven ($Q < 0.10$; Table 5). Of these, three significant SNP markers were associated with α-cellulose content (SNPs 18, 49 and 75);
one SNP marker each was associated with holocellulose, fiber length, fiber width and H traits were observed in the linkage population, with the marker effects that varied from 1.5% to 3.6% (Q < 0.10; Table 5).

SNPs 18 and 49 with α-cellulose, and SNP59 with fiber width were identified in both association and linkage populations (Tables 3 and 5). Because SNP49 was heterozygous in both parents (AC : AC) of the family-based linkage population, the effects of different genotype classes (AA, AC and CC) at SNP49 for α-cellulose content were similar in both populations (Figure 4), which are supported by the observation that SNP49 has an additive effect on cellulose content (Table 3 and Figure 4). Similarly, for the noncoding marker SNP18, the heterozygous trees (AT) showed higher average α-cellulose content than the homozygous trees TT (40.18% in AT and 39.20% in TT, they are significant), indicating that the minor allele T was dominant, and genotypic effects on α-cellulose content were consistent in both populations (Figure 4). In the linkage population, the SNP59 genotype was different in the parents (AA: AC) and the significant effects among the corresponding genotypic classes in SNP59 (20.29 μm in AA, 21.40 μm in AC) were consistent in the association population (Tables 3 and 5). Further analysis of the data suggested that these single markers also supported the haplotype-based associations with corresponding traits (Tables 4 and 5).

Transcript analysis of SNP genotypes

To determine whether these significant allelic SNPs affect the PtoCesA4 RNA transcript abundance, transcript levels were compared among the different genotypic classes for the ten significant SNPs (Q < 0.10, Table 3) in the association population and seven (Q < 0.10,
Table 5) in the linkage population, using RT–qPCR with gene-specific primers. Measurement of differential transcript abundance across three or two genotypic classes (10 trees for each genotype) for each of the seventeen SNPs, indicated that two markers (SNP41 and 49) exhibited significant differences in the RNA transcript levels among the three genotypes in the association population, but only SNP49 was detected in the linkage population (Figure 6). The genotypic abundance ratio estimates for these SNPs had very low standard errors, suggesting that these estimates are robust (Figure 6).

In the association population, for the marker SNP49 (exonic) associated with three traits, the highest transcript abundance was found in the AA group (0.6955), followed by the AC group (0.4307), and the transcript levels of the CC group were lowest (0.3308) (Figure 6-A). In examining genotype-specific transcript levels for SNP41 (5'UTR), the heterozygous trees (0.7511 in CT) for this marker showed higher relative abundance than the homozygous trees, and the transcript level differences between SNP homozygous trees were not significant (0.5533 in CC and 0.5606 in TT) (Figure 6B). In the linkage population, SNP 49 also showed a significant difference in transcript level among three distinct genotype classes, and the transcript levels for the groups AA, AC, and CC were 0.7819, 0.6322, and 0.4010, respectively (Figure 6-A). These differences in transcript abundance among genotypic classes are discussed below with their putative function and relationship with variation in phenotypic trait.

DISCUSSION
**PtoCesA4**, a highly targeted candidate gene belonging to key pathways of secondary cell wall biosynthesis was investigated using a more comprehensive approach than previous studies in trees (Wegrzyn et al. 2010; Sexton et al. 2011; Beaulieu et al. 2011; Dillon et al. 2010, 2012; Guerra et al. 2012). Phylogenetic analysis of *PtoCesA4* and other eukaryotic CesAs in land plant lineages revealed that *PtoCesA4* is an ortholog of *AtCesA4* and *PtiCesA4*. Functional studies of these putative orthologs informed the structure of *PtoCesA4* and its allelic diversity affecting complex traits controlling growth and lignocellulosic biosynthesis in *P. tomentosa*. Given the rapid decay of within-gene linkage disequilibrium in *Populus*, it is better to extend SNP discovery across the full-length gene sequence, including promoter regions, to avoid missing important allelic polymorphisms. Several small-effect single-SNP or haplotype-based associations were detected in this species, indicating wood properties are quantitative traits controlled by multiple alleles, and limited statistical power for single-marker or haplotype-based association method (Beaulieu et al. 2011; Du et al. 2013). Therefore, using linkage-LD mapping approaches combining with transcriptomic comparison of genotypic classes of each significant SNP, allowed us to examine functional *PtoCesA4* allelic variation responsible for complex quantitative traits related to lignocellulosic biosynthesis in trees.

**Nucleotide diversity and linkage disequilibrium in PtoCesA4**

Detailed knowledge of levels of nucleotide diversity and the extent of linkage disequilibrium in natural populations is important for understanding the forces responsible for evolutionary change, and for evaluating the precision and power of association mapping
As a prerequisite for SNP-based association mapping, a comprehensive investigation of the patterns of SNP distribution and frequency within the full-length *PtoCesA4* locus, and among natural populations of *P. tomentosa*, was required. Levels of average nucleotide diversity in coding regions were substantially lower than in the noncoding regions (Table 1), reflecting that the coding regions are conserved relative to the other regions under natural pressure. Within coding regions, the $\frac{\pi_{\text{nonsyn}}}{\pi_{\text{syn}}}$ ratio (0.0604) was significantly less than 1 for *PtoCesA4*, which was commonly observed in natural populations of forest trees (Krutovsky and Neale 2005; González-Martínez *et al.* 2006; Ingvarsson 2008a). Synonymous mutations occurring during evolution may be fixed with a higher probability than neutral ones due to purifying selection (Zhang *et al*. 2010b). Furthermore, an excess of transitional over transversional substitutions was found in this gene. There is a universal bias in favor of transitions over transversions, possibly as a result of the underlying chemistry of mutation, such as the relatively high rate of mutation of methylated cytosines to thymine, or particularly selection for codon-usage bias in coding regions (Keller *et al*. 2007; Ingvarsson 2008b). For synonymous polymorphism (47) in *PtoCesA4*, 80.1% (39) were transition substitutions, indicating that translational selection has shaped synonymous codon usage. This finding was supported by the most common model of synonymous codon usage that one or a few synonymous codons are preferentially used in genes with high codon bias, and such codons usually end in C or G (Ingvarsson 2008b). Moreover, the intron region harbored significantly less nucleotide diversity than synonymous sites (Table 1), similar to patterns recently reported for *P. balsamifera* (Olson *et al*. 2010), and *Medicago truncatula* (Branca *et al*. 2011). Introns experience higher
selective constraints than synonymous coding sites, possibly because introns harbor key regions for regulation of expression (Wray et al. 2003; Thumma et al. 2005). Surveys of SNPs diversity in gene have been mainly focused on exons, introns, and UTRs, with less attention to promoter regions. It will be of interest to compare the level of nucleotide diversity within the promoters with a much larger diverse survey of the other gene regions. In this study, a significantly higher frequency of polymorphisms was found in the promoter region than the other regions in PtoCesA4 (πT = 0.0230; Table 1). This finding suggests that this regulatory region may be relatively unstable and a hotspot for genetic change (Weickert et al. 2012).

In this study, the level of LD decay in PtoCesA4 was analyzed separately within each of the three climatic regions and for the complete natural population (Figure 3), and results showed that NW and S regions seem to have experienced similar histories. The NE region had a higher LD than the NW and S regions, consistent with the higher frequency of exclusive SNPs observed in this region (Table 2). The fine-scale LD pattern among polymorphic SNP markers in candidate genes may be influenced by gene conversion in different sampling populations. Generally, low LD might result from a species-wide scale of sampling, which incorporates the entire history of polymorphism and recombination over thousands of generations (Morrell et al. 2005; Kim et al. 2007). Our results support this sample-scale explanation, which show that the LD in PtoCesA4 within three climatic regions may be more extensive than the LD found in our range-wide P. tomentosa samples (Figure 3), consistent with previous studies (Olsen et al. 2010; Branca et al. 2011). However, a recent genome-wide study of the extensive linkage disequilibrium in P.
trichocarpa \( (r^2 > 0.2, \) within 3–6 kb), suggests that genome-wide association studies and genomic selection in natural populations may be more feasible in Populus than previously assumed (Slavov et al. 2012). Therefore, our future work will focus on estimation of LD decay with greater genomic coverage and exploration of the variability of haplotype structure across the entire genome. Such studies also will help to elucidate how Populus managed to adapt to a wide variety of environmental conditions (Ingvarsson and Street 2010).

Dissecting allelic polymorphisms underlying growth and wood properties

Estimating population structure is an important prerequisite in LD-based association analysis, and it is important to avoid false positives or spurious associations and to constrain association studies in association populations (Du et al., 2012). Mixed linear model (MLM) methods have proven useful in controlling for population structure and individual relatedness within association mapping studies, with the population structure matrix \( Q \) and the relative kinship matrix \( K \) as the covariance. In this study, Observed effect of the population structure (when \( K = 11 \) or \( K = 3 \), Du et al., 2012) on phenotypic variation is not significant, and the number and power of significant associations identified in two settings are stable (Data not shown), suggesting that tree species are ideal for the fine mapping of candidate genes and functional analysis of gene variants, as they are predominantly outcrossing, and have large, effective, relatively unstructured population sizes.

The variation in quantity and quality of primary and secondary wall cellulose in plants
is suggested to be the result of enzymatic activities of different types of cellulose synthase
(CesA) (Somerville 2006). Secondary cell walls have a higher percentage of cellulose, a
higher degree of polymerization, and a higher crystallinity than xylem primary walls (Joshi
et al. 2004). Because of the importance of secondary walls in determining wood quality
traits, many researchers have focused on secondary cell wall CesAs. In this study,
_PtoCesA4_ was originally isolated from a developing xylem cDNA library of _P. tomentosa_,
and was found to have a xylem-specific expression patterns (Figure 2). Similarly, its
putative ortholog in Arabidopsis (_AtCesA4_) was specifically associated with secondary cell
wall development (Atanassov et al. 2009), and the ortholog in _P. trichocarpa_ (_PtiCesA4_)
also expressed in developing xylem tissue undergoing secondary wall thickening (Suzuki et
al. 2006). In addition to the direct coding of CesA subunit proteins, genetic evidence has
confirmed an effect of CesA on wood chemical properties, influencing
cellulose/hemicellulose content as well as lignin content and composition (Song et al. 2010;
Wegrzyn et al. 2010). Several candidate genes in other pathways are also involved in
synthesizing cellulose (Szyjanowicz et al. 2004; Coleman et al. 2009). On the basis of
these studies, we dissected allelic polymorphisms within _PtoCesA4_, underlying growth and
wood properties, by using LD-based association in _P. tomentosa_, combined with
single-SNP linkage analysis. Due to the low LD in _P. tomentosa_ (Figure 3), once a
marker–trait association has been discovered and validated, it is likely that such a marker is
located in close proximity to the causal polymorphisms or even the functional variant itself
(Neale and Kremer 2011). Two non-synonymous markers (SNPs 49 and 59), and a
noncoding marker (SNP18) were associated with the same traits in both the association and
linkage populations (Tables 3 and 5), confirming the value of an integrated approach for characterizing the genetic basis of wood traits. We also found that the population differentiation ($Fst = 0.075$, Table 3) for these significant SNPs was greater than that ($Fst = 0.028$) of all common SNPs identified in the association population, which is consistent with the report that putative functional SNPs in genes in etiologic pathways for CVD show greater population differentiation than non-functional SNPs (Kullo and Ding 2007).

Using single marker association, a nonsynonymous substitution in exon3 of PtoCesA4 (SNP49) was in strong association with multiple traits ($\alpha$-cellulose, lignin, and H), and the modes of gene action appeared to be additive with the minor allele (A) conferring a lower lignin content and higher values in $\alpha$-cellulose and H. This marker also was identified with these three traits in the multi-SNP analysis (Table S3). This is consistent with the significant phenotypic correlation between these three traits (Du et al. 2013), and also represents a pleiotropic effect of PtoCesA4 on certain traits. Wood is composed of cellulose microfibrils embedded in a lignin-hemicellulose matrix. The observed associations of PtoCesA4 with diverse traits suggest that PtoCesA4 influences two distinct pathways (lignin and cellulose biosynthesis) in secondary cell wall synthesis (Song et al. 2010; Du et al. 2013). A similar phenomenon has been identified in previous studies (Thumma et al. 2009; Wegrzyn et al. 2010). Cellulose biosynthesis is co-expressed with other biological processes in plant vascular development, and the genes involved in these shared pathways often are functional homologs (Somerville et al. 2004; Eckert et al. 2012). For instance, genes encoding lignin monomer-polymerizing laccases and lignin monomer synthesis enzymes are among the most closely co-expressed genes with secondary cell wall AtCesA4,
7, and 8 (Persson et al. 2005). The inverse genotype effects in SNP49 between α-cellulose and lignin content might be indirectly related to carbon distribution toward the synthesis of C5 or C6 sugars (Guerra et al. 2012), which is in accordance with the significant negative phenotypic correlation between α-cellulose and lignin content in both P. tomentosa association and linkage populations (Du et al. 2013). Moreover, common haplotypes (SNPs 48-50) associated with α-cellulose traits surround SNP49, and this locus was identified in the multi-SNP associations (Table 4 and S3). We also observed a significant association between SNP49 and α-cellulose in the linkage population, with the same genotypic effect for this locus in association populations (Table 5 and Figure 4), suggesting that SNP49 may be a functional polymorphism that is in or near a locus involved in the control of α-cellulose content. This conjecture was also supported by the significant differences in expression among three genotype classes of SNP49 in either association or linkage populations (Figure 6-A).

Fibers are the most abundant secondary wall–containing cells in wood of dicot species. During secondary wall formation, highly coordinated expression of multiple genes controls cell elongation or secondary wall thickening of fibers (Burton et al. 2004; Zhong et al. 2006). For example, AtCesA7/IRX3 and AtCOBL4/IRX6 are coexpressed in tissues during secondary cell wall development, and loss-of-function mutation of either of these genes causes diminished cellulose content and loss of mechanical strength of the plant body (Brown et al. 2005). A mutant allele of AtCesA7 in fra5 (fragile fiber 5) causes a severe decrease in cellulose content and the thickness of fibers (Zhong et al. 2003). Cellulose is a biopolymer that provides a major contribution to secondary cell wall formation during cell
expansion and elongation (Xie et al. 2011). These early studies laid the research foundation for elucidation of a significant nonsynonymous association (SNP59) in exon 6 of PtoCesA4 with fiber width in *P. tomentosa* by using linkage-LD mapping, which demonstrated modes of gene action consistent with additive effects (Tables 3, 4, 5 and S3). Furthermore, a haplotype-based association with fiber width (SNPs 57-59) suggests that this locus may be located close to causative polymorphisms. This is consistent with the finding that nonsynonymous mutations play special roles in assigning functions to specific domains or motifs of the CESA (Zhong et al. 2003). Further analyzing the protein structure encoded by *PtoCesA4*, we found that the nonsynonymous mutation of amino acid 245 (Ser to Tyr) is close to the two putative transmembrane domains (TMDs) at the N-terminus (217-238 and 250-267), which are involved in CESAs protein–protein interactions (Joshi et al. 2004), suggesting that this nonsynonymous locus may affect the TMDs and also affect regulation of gene expression related to fiber width. CESAs are membrane-spanning proteins and small side chain residues often occur at the TMDs as a requirement of helix folding and structural stability. Zhang et al. (2009) identified a missense mutation (G858R) in the fifth TMD of the rice ortholog of *PtoCesA4* (OsCesA4); this mutation affects protein abundance in the plasma membrane and results in abnormal cell wall biosynthesis. Additionally, novel point mutations in the TMD have also been reported to affect cellulose synthesis in *Arabidopsis* (Chen et al. 2005).

Many functional analyses of SNPs have examined coding regions and splicing sites in candidate genes related to wood traits that can alter proteins and mRNA splicing. However, SNPs in non-coding regulatory regions can also influence important biological regulation.
(Thumma et al. 2009; Beaulieu et al. 2011). d’Alesio et al. (2005) detected that several SNPs are predicted to be related to genes by influencing the binding affinity of transcription factors in the promoter region. In this work, we detected a significant marker (SNP 18) at 273 bp (T/A) upstream of the transcriptional start site of the \textit{PtoCesA4} promoter (Figure 1). Genotypic effect analysis of \(\alpha\)-cellulose content in either association or linkage population showed that the trees heterozygous (AT) for this marker showed higher average \(\alpha\)-cellulose content than the homozygous trees (Figure 4), indicating overdominance. Moreover, a common haplotype (SNPs 16-18) associated with \(\alpha\)-cellulose traits, and SNP18 was also determined by using multi-SNPs association model (Table 4 and S3). These results support that SNP18 maybe have a regulatory effect on \textit{PtoCesA4} expression or it could to be in very strong LD with a nearby regulatory polymorphism, the detailed regulatory mechanisms of this locus will require further investigation. Although the mutation in the 5' flanking region did not result in an amino acid changes, phenotypic traits can be affected because 5'UTRs play crucial roles in the regulation of gene expression, especially for transcriptional, mRNA stability, translational efficiency, or sub-cellular localization (Miyamoto et al. 2007; Lin et al. 2012). In our study, SNP41, located in the 5' UTR of \textit{PtoCesA4}, had a significant association (Q < 0.10) with \(\alpha\)-cellulose in the association populations (Table 3). The estimated allelic effects of SNP 41 on \(\alpha\)-cellulose corresponded well with estimates of transcript levels (Figure 6-B); \textit{i.e.}, heterozygotes (CT) had both higher average \(\alpha\)-cellulose content and higher transcript levels than the homozygotes (CC and TT), suggesting that SNP41 may be a functional polymorphism affecting the regulation of gene expression. However, our follow-up study of this SNP in
the linkage population did not support this observation. Similarly, no replication case has been reported in several previous association studies on wood traits (Dillon et al. 2010, 2012; Du et al. 2013). The differences between association and linkage populations may explain the “lack of validation” for this association, including their genetic background, complex gene-environment interactions, mapping resolution, population structure, and age-dependent effects (Du et al. 2013). In this study, the linkage population has a limited genomic data set from both the parents and the interspecific genetic background of the female parent. Therefore, improved power to detect and validate associations in future experiments could be achieved by establishing validation populations with families, or clonal material, from the discovery population of the same species (Dillon et al. 2010).

Conclusions

Tissue-specific expression profiles revealed that *PtoCesA4* is highly expressed during secondary cell wall formation. Therefore, selection of optimal candidate genes through different approaches such as EST database searches, transcript abundance profiles, QTL mapping and comparison of orthologs in model or related species, is very important for identifying useful alleles located within functional genes controlling traits of interest (Neale and Savolainen 2004; Thumma et al. 2005). Our work revealed that the greater length may lead to slightly higher LD than candidate genes analyzed previously in *Populus*; and combined with a recent genome-wide study of linkage disequilibrium in *P. trichocarpa* (Slavov et al. 2012), suggesting that LD studies in *P. tomentosa* should focus on a better understanding of the variability of haplotype structure across the entire genome.
Comparatively, the preliminary application of multi-SNP analysis in PtoCesA4 suggests that it will be promising to conduct association studies with virtually all the related genes that share biological pathways, and to have a more complete understanding of the genetic architecture of quantitative variation (Eckert et al. 2012; Resende et al. 2012).

Wood-quality traits are quantitative traits controlled by multiple genes, with a moderate to high degree of heritability; but growth traits have relatively low heritability compared to wood-property traits (Thumma et al., 2010). Therefore, several significant SNP associations with wood traits, obtained using linkage-LD mapping approaches combined with RNA transcript abundance among the genotypes of each significant SNP, represent important progress toward the identification of allelic variation responsible for wood traits and the development of successful marker-aided selection in trees. In the coming years, the rapid development of high-throughput sequencing is very likely to drive association studies towards genome-wide studies in trees.

Acknowledgments

We thank Profs Ronald R. Sederoff, Zhao-Bang Zeng, and Ross W. Whetten (NC State University, Raleigh, NC, USA), Harry Wu (Umeå Plant Science Centre, Swedish University of Agricultural Sciences, Umeå, Sweden), and Rongling Wu (Center for Statistical Genetics, Pennsylvania State University, Pennsylvania, USA) for their detailed comments and specific suggestions for improving the manuscript. This work was supported by the grants from the Introduction of Foreign Advanced Agricultural Science and Technology into China (No. 2009-4-22), the State Key Basic Research Program of China.
(No. 2012CB114506), and the Project of the National Natural Science Foundation of China (No. 31170622, 30872042).

LITERATURE CITED


Beaulieu, J., T. Doerksen, B. Boyle, S. Clement, M. Deslauriers et al., 2011 Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. Genetics 188: 197-214.


Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss et al., 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics


Dillon S. K., M. Nolan, W. Li, C. Bell, H. X. Wu et al., 2010 Allelic variation in cell wall candidate genes affecting solid wood properties in association populations and land races of *Pinus radiata*. Genetics 185: 1477-1487.


Huang, Z. H., 1992 The study on the climatic regionalization of the distributional region of 

Ingvarsson, P. K., 2008a Multilocus patterns of nucleotide polymorphism and the 

Ingvarsson, P. K., 2008b Molecular evolution of synonymous codon usage in *Populus*. 

Ingvarsson, P. K., and N. R. Street, 2011 Association genetics of complex traits in plants. 

Joshi, C. P., S. Bhandari, P. Ranjan, U. C. Kalluri, X. Liang *et al.*, 2004 Genomics of 

Keller, I., D. Bensasson, and R. A. Nichols, 2007 Transition-transversion bias is not 
doi:10.1371/journal.pgen.0030022.

Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark *et al.*, 2007 Recombination and 

Krutovsky, K. V., and D. B. Neale, 2005 Nucleotide diversity and linkage disequilibrium in 
cold-hardiness-and wood quality-related candidate genes in Douglas Fir. Genetics 171:  
2029-2041.

Kullo, I. J., and K. Ding, 2007 Patterns of population differentiation of candidate genes for 

Kumar, M., S. Thammannagowda, V. Bulone, V. Chiang, K. H. Han *et al.*, 2009 An update 
on the nomenclature for the cellulose synthase genes in *Populus*. Trends Plant Sci. 14:


Myles, S., J. Peiffer, P. J. Brown, E. S. Ersoz, Z. Zhang *et al.*, 2009 Association mapping: Critical considerations shift from genotyping to experimental design. Plant Cell 21:


Resende, M. F., P. Muñoz, M. D. Resende, D. J. Garrick, R. L. Fernando et al., 2012


Szyjanowicz, P. M., I. McKinnon, N. G. Taylor, J. Gardiner, M. C. Jarvis et al., 2004 The irregular xylem 2 mutant is an allele of korrigan that affects the secondary cell wall of Arabidopsis thaliana. Plant J. 37: 730-740.


Weickert, C. S., Y. Tiwari, P. R. Schofield, B. J. Mowry, and J. M. Fullerton, 2012 Schizophrenia-associated HapICE haplotype is associated with increased NRG1 type
III expression and high nucleotide diversity. Transl. Psychiatry 2: e104

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer et al., 2003 The evolution of

Wu, L., C. P. Joshi, and V. L. Chiang, 2000 A xylem-specific cellulose synthase gene from

Xie, L., C. Yang, and X. Wang, 2011 Brassinosteroids can regulate cellulose biosynthesis
by controlling the expression of CESA genes in Arabidopsis. J. Exp. Bot. 62:
4495-4506.

method for association mapping that accounts for multiple levels of relatedness. Nat.

Zhang, B., L. Deng, Q. Qian, G. Xiong, D. Zeng et al., 2009 A missense mutation in the
transmembrane domain of CESA4 affects protein abundance in the plasma membrane
and results in abnormal cell wall biosynthesis in rice. Plant Mol. Biol. 71: 509-524.

Genomics. 284: 105-119.

Zhang, D. Q., X. H. Yang, Z. Y. Zhang, and B. Li, 2010b Expression and nucleotide

Zhong, R., T. Demura, and Z. H. Ye, 2006 SND1, a NAC domain transcription factor, is a
key regulator of secondary wall synthesis in fibers of Arabidopsis. Plant Cell 18:
Table 1 Nucleotide polymorphism at the PtoCesA4 locus

<table>
<thead>
<tr>
<th>Region</th>
<th>Length (bp)</th>
<th>No. of polymorphic sites</th>
<th>Frequency (bp⁻¹)</th>
<th>Transitions/transversions</th>
<th>Nucleotide diversity</th>
<th>π</th>
<th>θw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter</td>
<td>1111</td>
<td>86</td>
<td>13</td>
<td>1.676</td>
<td>0.0230</td>
<td>0.0246</td>
<td></td>
</tr>
<tr>
<td>5'UTR</td>
<td>297</td>
<td>9</td>
<td>33</td>
<td>3.500</td>
<td>0.0053</td>
<td>0.0104</td>
<td></td>
</tr>
<tr>
<td>Exon1</td>
<td>51</td>
<td>0</td>
<td>/</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Intron1</td>
<td>93</td>
<td>5</td>
<td>19</td>
<td>1.500</td>
<td>0.0060</td>
<td>0.0126</td>
<td></td>
</tr>
<tr>
<td>Exon2</td>
<td>202</td>
<td>6</td>
<td>34</td>
<td>0.500</td>
<td>0.0097</td>
<td>0.0105</td>
<td></td>
</tr>
<tr>
<td>Intron2</td>
<td>82</td>
<td>6</td>
<td>14</td>
<td>1.000</td>
<td>0.0083</td>
<td>0.0172</td>
<td></td>
</tr>
<tr>
<td>Exon3</td>
<td>125</td>
<td>2</td>
<td>63</td>
<td>1.000</td>
<td>0.0047</td>
<td>0.0056</td>
<td></td>
</tr>
<tr>
<td>Intron3</td>
<td>299</td>
<td>10</td>
<td>30</td>
<td>1.500</td>
<td>0.0058</td>
<td>0.0080</td>
<td></td>
</tr>
<tr>
<td>Exon4</td>
<td>67</td>
<td>1</td>
<td>67</td>
<td>/</td>
<td>0.0072</td>
<td>0.0035</td>
<td></td>
</tr>
<tr>
<td>Intron4</td>
<td>96</td>
<td>3</td>
<td>32</td>
<td>/</td>
<td>0.0016</td>
<td>0.0074</td>
<td></td>
</tr>
<tr>
<td>Exon5</td>
<td>151</td>
<td>1</td>
<td>151</td>
<td>/</td>
<td>0.0033</td>
<td>0.0016</td>
<td></td>
</tr>
<tr>
<td>Intron5</td>
<td>79</td>
<td>6</td>
<td>13</td>
<td>5.000</td>
<td>0.0160</td>
<td>0.0179</td>
<td></td>
</tr>
<tr>
<td>Exon6</td>
<td>613</td>
<td>11</td>
<td>56</td>
<td>4.500</td>
<td>0.0042</td>
<td>0.0058</td>
<td></td>
</tr>
<tr>
<td>Intron6</td>
<td>173</td>
<td>9</td>
<td>19</td>
<td>8.000</td>
<td>0.0075</td>
<td>0.0164</td>
<td></td>
</tr>
<tr>
<td>Exon7</td>
<td>138</td>
<td>3</td>
<td>46</td>
<td>2.000</td>
<td>0.0041</td>
<td>0.0051</td>
<td></td>
</tr>
<tr>
<td>Intron7</td>
<td>92</td>
<td>0</td>
<td>/</td>
<td>/</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Exon8</td>
<td>126</td>
<td>4</td>
<td>32</td>
<td>3.000</td>
<td>0.0016</td>
<td>0.0075</td>
<td></td>
</tr>
<tr>
<td>Intron8</td>
<td>88</td>
<td>4</td>
<td>22</td>
<td>/</td>
<td>0.0028</td>
<td>0.0134</td>
<td></td>
</tr>
<tr>
<td>Exon9</td>
<td>213</td>
<td>5</td>
<td>43</td>
<td>1.500</td>
<td>0.0024</td>
<td>0.0055</td>
<td></td>
</tr>
<tr>
<td>Intron9</td>
<td>125</td>
<td>4</td>
<td>31</td>
<td>0.333</td>
<td>0.0061</td>
<td>0.0076</td>
<td></td>
</tr>
<tr>
<td>Exon10</td>
<td>510</td>
<td>11</td>
<td>46</td>
<td>10.000</td>
<td>0.0065</td>
<td>0.0065</td>
<td></td>
</tr>
<tr>
<td>Intron10</td>
<td>294</td>
<td>8</td>
<td>37</td>
<td>1.667</td>
<td>0.0084</td>
<td>0.0059</td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td>Length</td>
<td>Ns</td>
<td>Nc</td>
<td>Nv</td>
<td>P1</td>
<td>P2</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>Exon11</td>
<td>351</td>
<td>4</td>
<td>88</td>
<td>3.000</td>
<td>0.0008</td>
<td>0.0027</td>
<td></td>
</tr>
<tr>
<td>Intron11</td>
<td>132</td>
<td>5</td>
<td>26</td>
<td>4.000</td>
<td>0.0086</td>
<td>0.0089</td>
<td></td>
</tr>
<tr>
<td>Exon12</td>
<td>582</td>
<td>7</td>
<td>83</td>
<td>6.000</td>
<td>0.0031</td>
<td>0.0036</td>
<td></td>
</tr>
<tr>
<td>3'UTR</td>
<td>331</td>
<td>8</td>
<td>41</td>
<td>1.667</td>
<td>0.0064</td>
<td>0.0057</td>
<td></td>
</tr>
<tr>
<td>Total silent</td>
<td>3856.32</td>
<td>210</td>
<td>18</td>
<td>2.060</td>
<td>0.0124</td>
<td>0.0152</td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td>714.32</td>
<td>47</td>
<td>15</td>
<td>4.875</td>
<td>0.0149</td>
<td>0.0194</td>
<td></td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>2411.68</td>
<td>8</td>
<td>301</td>
<td>3.000</td>
<td>0.0009</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td>Total exon</td>
<td>3129</td>
<td>55</td>
<td>57</td>
<td>2.500</td>
<td>0.0040</td>
<td>0.0048</td>
<td></td>
</tr>
<tr>
<td>Total intron</td>
<td>1553</td>
<td>60</td>
<td>26</td>
<td>2.588</td>
<td>0.0065</td>
<td>0.0105</td>
<td></td>
</tr>
<tr>
<td>Totalb</td>
<td>6421</td>
<td>218</td>
<td>29</td>
<td>2.169</td>
<td>0.0080</td>
<td>0.0098</td>
<td></td>
</tr>
</tbody>
</table>

1 Regions containing indels are excluded from the calculation;  
*Total silent = synonymous plus noncoding sites;  
*Total = silent sites plus nonsynonymous sites
Table 2  Summary of nucleotide variations for PtoCesA4 in *Populus tomentosa* natural populations from three climatic regions.

<table>
<thead>
<tr>
<th>Population</th>
<th>N</th>
<th>S</th>
<th>S_{al}</th>
<th>π_{tot}</th>
<th>π_{sil}</th>
<th>π_{s}</th>
<th>π_{n}</th>
<th>Tajima (1989)'s D</th>
<th>Fu and Li (1993)'s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeastern</td>
<td>14</td>
<td>166</td>
<td>26</td>
<td>0.0074</td>
<td>0.0114</td>
<td>0.0130</td>
<td>0.0008</td>
<td>-0.5013</td>
<td>-0.4495</td>
</tr>
<tr>
<td>Southern region</td>
<td>13</td>
<td>149</td>
<td>5</td>
<td>0.0085</td>
<td>0.0131</td>
<td>0.0166</td>
<td>0.0010</td>
<td>0.4973</td>
<td>-0.1296</td>
</tr>
<tr>
<td>Northwestern</td>
<td>13</td>
<td>144</td>
<td>5</td>
<td>0.0082</td>
<td>0.0127</td>
<td>0.0141</td>
<td>0.0010</td>
<td>0.5355</td>
<td>0.0091</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>218</td>
<td>/</td>
<td>0.0080</td>
<td>0.0124</td>
<td>0.0149</td>
<td>0.0009</td>
<td>-0.6498</td>
<td>-2.2043</td>
</tr>
</tbody>
</table>

N=Number of sequences sampled; S=Number of segregating sites; S_{al}=Polymorphic exclusive biallelic mutations in the studied group; π_{tot}=Average nucleotide diversity in full gene; π_{sil}= Average nucleotide diversity in synonymous and noncoding sites, π_{s}= Average nucleotide diversity of synonymous mutation π_{n}= Average nucleotide diversity of non-synonymous mutation
Table 3 Summary of significant SNP marker-trait pairs identified in the *Populus tomentosa* association population using the mixed linear model (MLM) after a correction for multiple testing.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Locus</th>
<th>Position</th>
<th>Mutation</th>
<th>Fst</th>
<th>Association population (N = 460)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P-value</td>
</tr>
<tr>
<td>Lignin</td>
<td>SNP44</td>
<td>Intron1</td>
<td>[G : T]</td>
<td>0.057</td>
<td>0.0012</td>
</tr>
<tr>
<td></td>
<td>SNP49</td>
<td>Exon 3</td>
<td>[C : A]⁵</td>
<td>0.114</td>
<td>0.0025</td>
</tr>
<tr>
<td>α-cellulose</td>
<td>SNP3</td>
<td>Promoter</td>
<td>[G : A]</td>
<td>0.077</td>
<td>0.0015</td>
</tr>
<tr>
<td></td>
<td>SNP18</td>
<td>Promoter</td>
<td>[A : T]</td>
<td>0.039</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td>SNP41</td>
<td>5’UTR</td>
<td>[C : T]</td>
<td>0.100</td>
<td>3.02E-05</td>
</tr>
<tr>
<td></td>
<td>SNP49</td>
<td>Exon 3</td>
<td>[C : A]⁶</td>
<td>0.114</td>
<td>0.0031</td>
</tr>
<tr>
<td>Holocellulose</td>
<td>SNP45</td>
<td>Exon 2</td>
<td>[C : A]</td>
<td>0.050</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>SNP81</td>
<td>Intron 10</td>
<td>[T : C]</td>
<td>0.052</td>
<td>0.0002</td>
</tr>
<tr>
<td>Fiber width</td>
<td>SNP59</td>
<td>Exon 6</td>
<td>[A : C]</td>
<td>0.130</td>
<td>0.0008</td>
</tr>
<tr>
<td>Diameter at breast</td>
<td>SNP48</td>
<td>Intron 2</td>
<td>[A : T]</td>
<td>0.091</td>
<td>0.0009</td>
</tr>
<tr>
<td>height (D)</td>
<td>SNP75</td>
<td>Exon 10</td>
<td>[T : C]</td>
<td>0.036</td>
<td>3.15E-05</td>
</tr>
<tr>
<td></td>
<td>SNP81</td>
<td>Intron 10</td>
<td>[T : C]</td>
<td>0.052</td>
<td>0.0003</td>
</tr>
<tr>
<td>Tree height (H)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNP</td>
<td>Exon</td>
<td>Mutation</td>
<td>ns</td>
<td>s</td>
<td>Fst</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>-------------------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>SNP49</td>
<td>3</td>
<td>[C : A]³ns</td>
<td>0.114</td>
<td>0.0012</td>
<td>0.0551</td>
</tr>
<tr>
<td>SNP75</td>
<td>10</td>
<td>[T : C]¹s</td>
<td>0.036</td>
<td>3.02E-05</td>
<td>0.0035</td>
</tr>
</tbody>
</table>

ns, nonsynonymous polymorphism; s, synonymous polymorphism; Fst = Variation due to differentiation among subpopulations $R^2 =$ percentage of the phenotypic variance explained; $P$-value = significance level for association (significance is $P \leq 0.05$); $Q$-value = a correction for multiple testing (false discovery rate FDR ($Q$) ≤ 0.10).

¹Calculated as the difference between the phenotypic means observed within each homozygous class ($2a=|G_{BB}-G_{bb}|$, where $G_{ij}$ is the trait mean in the $ij$th genotypic class).

²Calculated as the difference between the phenotypic mean observed within the heterozygous class and the average phenotypic mean across both homozygous classes [$d=G_{Bb}-0.5(G_{BB}+G_{bb})$, where $G_{ij}$ is the trait mean in the $ij$th genotypic class].

³$σ_p$, standard deviation for the phenotypic trait under consideration.

⁴Allele frequency of either the derived or minor allele. Single nucleotide polymorphism (SNP) alleles corresponding to the frequency listed are given in parentheses.

⁵The additive effect was calculated as $a=p_b(G_{BB}) + p_b(G_{Bb})-G$, where $G$ is the overall trait mean, $G_{ij}$ is the trait mean in the $ij$th genotypic class and $p_i$ is the frequency of the $i$th marker allele. These values were always calculated with respect to the minor allele.
Table 4 List of haplotypes with significant associations with wood quality and growth traits in the *P. tomentosa* association population (n = 460) after a correction for multiple testing.

<table>
<thead>
<tr>
<th>Trait</th>
<th>P-value</th>
<th>Q-value</th>
<th>$R^2$(%)</th>
<th>Significant haplotypes</th>
<th>Haplotype frequency</th>
<th>Single-marker associations$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lignin</td>
<td>0.0012</td>
<td>0.0487</td>
<td>3.4</td>
<td>SNPs 2-4</td>
<td></td>
<td>/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T-G-T</td>
<td>0.28</td>
<td>SNP44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G-A-C</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.0052</td>
<td>0.0760</td>
<td>3.7</td>
<td>SNPs 44-46</td>
<td></td>
<td>(Lignin, $Q = 0.0551$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G-A-T</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T-A-T</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>α-cellulose</td>
<td>0.0019</td>
<td>0.0532</td>
<td>3.8</td>
<td>SNPs 1-3</td>
<td></td>
<td>SNP3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T-G-A</td>
<td>0.25</td>
<td>(α-cellulose, $Q = 0.0629$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T-T-G</td>
<td>0.43</td>
<td>SNP18</td>
</tr>
<tr>
<td></td>
<td>0.0015</td>
<td>0.0487</td>
<td>2.8</td>
<td>SNPs 16-18</td>
<td></td>
<td>(α-cellulose, $Q = 0.0551$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T-A-A</td>
<td>0.08</td>
<td>SNP49</td>
</tr>
<tr>
<td></td>
<td>0.0063</td>
<td>0.0922</td>
<td>5.6</td>
<td>SNPs 48-50</td>
<td></td>
<td>(α-cellulose, $Q = 0.0948$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T-A-C</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>T-C-A</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>Holocellulose</td>
<td>0.0040</td>
<td>0.0713</td>
<td>3.0</td>
<td>SNPs 38-40</td>
<td></td>
<td>/</td>
</tr>
<tr>
<td>SNP</td>
<td>Fiber length</td>
<td>Fiber width</td>
<td>D</td>
<td>V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>--------------</td>
<td>-------------</td>
<td>------</td>
<td>------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNPs 44-46</td>
<td>0.0023 0.0579 5.1</td>
<td>SNPs 89-91</td>
<td>0.0051 0.0760 4.0</td>
<td>SNPs 47-49</td>
<td>0.0005 0.0187 2.7</td>
<td>SNPs 81-83</td>
</tr>
<tr>
<td>T-G-A</td>
<td>T-A-T</td>
<td>SNP45</td>
<td>0.0023 0.0579 5.1</td>
<td>SNP59</td>
<td>C-T-A</td>
<td>SNP81</td>
</tr>
<tr>
<td>0.12</td>
<td>0.09</td>
<td>(Holocellulose, $Q = 0.0142$)</td>
<td></td>
<td>0.05</td>
<td>0.17</td>
<td>(D, $Q =$ 0.0454)</td>
</tr>
<tr>
<td>T-G-A</td>
<td>T-A-T</td>
<td>SNP45</td>
<td></td>
<td>0.29</td>
<td>0.17</td>
<td>(D, $Q =$ 0.0454)</td>
</tr>
<tr>
<td>G-C-C</td>
<td>G-C-C</td>
<td>SNP45</td>
<td></td>
<td>0.13</td>
<td>0.21</td>
<td>(D, $Q =$ 0.0454)</td>
</tr>
<tr>
<td>G-A-C</td>
<td>G-A-C</td>
<td>SNP45</td>
<td></td>
<td>0.20</td>
<td>0.17</td>
<td>(D, $Q =$ 0.0454)</td>
</tr>
</tbody>
</table>

D = Diameter at breast height; V = stem volume; $R^2$ = percentage of the phenotypic variance explained; P-value = the significant level for haplotype-based association (the significance is $P \leq 0.05$); $Q$-value = a correction for multiple testing (false discovery rate FDR ($Q \leq 0.10$)); $^1$ Significant single-marker associations with the lowest $Q$ value (FDR $Q \leq 0.10$) relating to the significant haplotype–trait association; /, no data was
identified in this study.
Table 5 Summary of significant SNP marker-trait pairs identified in *PtoCesA4*, using a linkage population, after correction for multiple testing errors.

<table>
<thead>
<tr>
<th>Trait</th>
<th>Locus</th>
<th>Position</th>
<th>Alleles of parents (Female : Male)</th>
<th>Linkage population (N = 1200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><em>P</em>-value (<em>P</em> ≤ 0.05)</td>
</tr>
<tr>
<td>α-cellulose</td>
<td>SNP18</td>
<td>Promoter</td>
<td>[TT : AT]</td>
<td>0.0036</td>
</tr>
<tr>
<td></td>
<td>SNP49</td>
<td>Exon 3</td>
<td>[AC : AC]</td>
<td>0.0015</td>
</tr>
<tr>
<td></td>
<td>SNP75</td>
<td>Exon 10</td>
<td>[CT : CT]</td>
<td>0.0019</td>
</tr>
<tr>
<td>Holocellulose</td>
<td>SNP88</td>
<td>Exon 12</td>
<td>[AG : AG]</td>
<td>0.0034</td>
</tr>
<tr>
<td>Fiber length</td>
<td>SNP70</td>
<td>Intron 9</td>
<td>[AT : AT]</td>
<td>0.0013</td>
</tr>
<tr>
<td>Fiber width</td>
<td>SNP59</td>
<td>Exon 6</td>
<td>[AA : AC]</td>
<td>0.0044</td>
</tr>
<tr>
<td>Tree height (H)</td>
<td>SNP51</td>
<td>Intron 3</td>
<td>[AC : AC]</td>
<td>2.55E-05</td>
</tr>
</tbody>
</table>

\(R^2\) = percentage of the phenotypic variance explained; *P*-value = significance level for association (significance is *P* ≤ 0.05); *Q*-value = a correction for multiple testing (false discovery rate FDR (*Q*) ≤ 0.10).
Figure 1. *PtoCesA4* gene structure and the positions of common SNPs (minor allele frequencies > 0.10). All common SNPs are represented by dark spots; putative transcription factor binding sites around SNPs in the *PtoCesA4* promoter were predicted, numbers above the promoter region indicate the positions of putative transcription factor binding sites in base pairs relative to the predicted transcription start site. A: zinc-binding domain, B-I: two transmembrane helices in the N-terminal region and six in the C-terminal region.
Figure 2. Relative transcript levels of *PtoCesA4* in *Populus tomentosa* tissues and organs. The error bars represent + SD.
Figure 3. The decay of short-range linkage disequilibrium within *PtoCesA4* for all samples and each climatic region. Pairwise correlations between SNPs are plotted against the physical distance between the SNPs in base pairs. The curves describe the nonlinear regressions of $r^2$ (Er2) onto the physical distance in base pairs.
Figure 4. Genotypic effects of the significant single nucleotide polymorphisms in *PtoCesA4* on the same phenotypic trait in association and linkage populations. The marker SNP49 in exon 3 of *PtoCesA4*, a non-synonymous mutation, which results in an encoded amino acid change from His to Asn, was significantly associated with \( \alpha \)-cellulose content in association and linkage populations. The AA homozygotes were associated with higher \( \alpha \)-cellulose values and CC homozygotes were associated with lower, and mean values in AC heterozygotes were medium in both populations, which is supported by the observation that SNP49 has an additive effect on gene action in...
The nonsynonymous marker SNP59 in exon 6 of *PtoCesA4* significantly associated with fiber width in both populations, and shows patterns of gene action consistent with additive effects on fiber width. The A allele at SNP59 causes a Ser to Tyr amino acid substitution. SNP18 from the promoter of *PtoCesA4* showed significant association with α-cellulose content in both populations. The differences in α-cellulose content among the three genotypes of this marker indicate that patterns of gene action are consistent with overdominance effects. ‘P1’ represents the female clone ‘YX01’ (*Populus alba × Populus glandulosa*), ‘P2’ represents the male clone ‘LM 50’ (*Populus tomentosa*) ‘F1’ represents the hybrid progeny.
Figure 5. Multilocus single nucleotide polymorphism (SNP) models explain a large percentage of the phenotypic variance for growth and wood properties in the *Populus tomentosa* association population. The gray line and points denote the numbers of SNPs identified for each trait and the marker effects ($R^2$) explained by the list of SNPs identified using the Bayesian mixed linear model in the Bayesian association with missing data (BAMD) program in R (http://cran.r-project.org/package=BAMD).
Figure 6. *PtoCesA4* transcript abundance varies among genotypic classes for significant SNP associations. (A) Transcript abundance variation of three genotypic classes for SNP49 in both association and linkage populations. The black and gray lines represent the transcript levels among three genotypic classes in association and linkage populations, respectively. (B) The relative mRNA transcript levels of *PtoCesA4* among three genotypic classes for SNP41, a significant noncoding marker in 5'UTR region of *PtoCesA4*. The error bars represent ± SD.