Bioenerg. Res. (2010) 3:172–182
DOI 10.1007/s12155-009-9060-z

Bioinformatics-Based Identification of Candidate Genes from QTLs Associated with Cell Wall Traits in *Populus*

Priya Ranjan · Tongming Yin · Xinye Zhang · Udaya C. Kalluri · Xiaohan Yang · Sara Jawdy · Gerald A. Tuskan

Published online: 6 November 2009

Abstract Quantitative trait locus (QTL) studies are an integral part of plant research and are used to characterize the genetic basis of phenotypic variation observed in structured populations and inform marker-assisted breeding efforts. These QTL intervals can span large physical regions on a chromosome comprising hundreds of genes, thereby hampering candidate gene identification. Genome history, evolution, and expression evidence can be used to narrow the genes in the interval to a smaller list that is manageable for detailed downstream functional genomics characterization. Our primary motivation for the present study was to address the need for a research methodology that identifies candidate genes within a broad QTL interval. Here we present a bioinformatics-based approach for subdividing candidate genes within QTL intervals into alternate groups of high probability candidates. Application of this approach in the context of studying cell wall traits, specifically lignin content and S/G ratios of stem and root in *Populus* plants, resulted in manageable sets of genes of both known and putative cell wall biosynthetic function. These results provide a roadmap for future experimental work leading to identification of new genes controlling cell wall recalcitrance and, ultimately, in the utility of plant biomass as an energy feedstock.

Electronic supplementary material The online version of this article (doi:10.1007/s12155-009-9060-z) contains supplementary material, which is available to authorized users.

P. Ranjan (✉) · T. Yin · X. Zhang · U. C. Kalluri · X. Yang · S. Jawdy · G. A. Tuskan
Environmental Sciences Division,
Oak Ridge National Laboratory,
Oak Ridge, TN 37831, USA
e-mail: ranjanp@ornl.gov

P. Ranjan · T. Yin · X. Zhang · U. C. Kalluri · X. Yang · S. Jawdy · G. A. Tuskan
The Bioenergy Science Center, Oak Ridge National Laboratory,
Oak Ridge, TN 37831, USA

Keywords *Populus* · Whole-genome duplication · Quantitative trait loci · Wood chemistry · Syringyl lignin · Guaiacyl lignin · Biofuels

Abbreviations

QTL	Quantitative trait loci
S/G	Syringyl and guaiacyl ratio
RL	Root lignin content
RSG	Root S/G ratio
SL	Stem lignin content
SSG	Stem S/G ratio
SSR	Simple sequence repeats

Introduction

Plant biomass has recently been promoted as a source of renewable feedstock for the conversion to liquid transportation fuels [13, 15, 22, 29]. Plant cell walls can be biochemically or thermochemically deconstructed into the primary subcomponents (cellulose, hemicellulose, and lignin) necessary for this conversion [14, 19, 20, 26]. The carbohydrate fractions are used as feedstocks for sugar and ultimately ethanol production, and lignins are typically separated and used in combustion processes to fuel the reactions. The resistance of lignin, an amorphic polymer, to separate from the carbohydrate fractions during the deconstruction phase has made lignin a target for overcoming recalcitrance [12].

Lignin, a complex polyphenolic polymer, is one of the most abundant polymers on earth. Lignin content of the cell wall influences the cell rigidity, drought tolerance, and insect and disease resistance [7]. The biochemical pathway for lignin biosynthesis is fairly well characterized and involves approximately 12–15 enzyme-regulated steps controlling the conversion of single aldehyde to syringyl

and guaiacyl precursors [5, 37]. Lignin content varies across the tissue types and organs of a plant with developmental age and environmental interactions [32]. These responses are genetically controlled and heritability for lignin is moderately high [20]. The rate limiting/critical steps in lignin formation are not yet fully determined though several studies have used reverse genetic approaches and expression analysis to modify and/or characterize lignin composition in transgenic plant materials [9, 10, 23, 24].

Lignin and other cell wall traits display a pattern of continuous phenotypic distribution rather than discrete, Mendelian distribution. Such traits are typically polygenic in nature and are influenced by the environment in which they occur. Genetic mapping can be used to compare the inheritance pattern of a trait and establish the chromosomal regions associated with such phenotypes. These chromosomal intervals may encompass one or more genes responsible for the trait and are known as quantitative trait loci (QTLs).

After the identification of QTL intervals, filtering the list of genes down to a subset of likely candidates is a difficult task. The length of the QTL intervals may be in mega base pairs (Mbp) and include hundreds of genes. One approach to reducing the number of candidate genes is to conduct further experiments using larger numbers of segregating progeny to reduce the QTL interval. Then, classical methods such as positional cloning [25, 27] and insertional mutagenesis [3, 30] can be used to identify influential genes. A complementary approach would be to use the bioinformatics tools and genome information to assign genes in the QTL interval to bins of higher probability than other candidates. This gives a smaller number of candidate genes that can be verified using transgenesis.

The recent availability of several draft and fully sequenced plant genomes have shed light on the evolutionary history of genome structure, and the role whole-genome duplication events have played in determining genome structure and gene family evolution. It is becoming apparent that nearly all plant genomes have experienced at least one whole-genome duplication event [18, 39]. These events have influenced gene family evolution and created opportunities for paralogous genes to experience neo-functionalization and/or sub-functionalization within all gene families [8, 28, 36].

The *Populus* genome contains three whole-genome duplication events [35]. The most recent, the Salicoid duplication, is found only in members of the Salicoid family and is present in approximately 8,000 paralogous gene pairs. The second duplication event, shared by *Populus* and *Arabidopsis*, is found in 3,500 paralogous gene pairs in *Populus*. In addition, the molecular clock in *Populus* is ticking a rate that is six times slower than in *Arabidopsis*, creating a duplicated molecular preservation of the ancestral genome within the extant *Populus* genome [35]. Together these genomic features can complicate genome assembly, annotation as well as map-based cloning of individual gene(s) responsible for specific phenotypes.

We use a combination of traditional QTL mapping, comparative intragenomic analysis, estimates of gene divergence, and differential expression evidence to identify regions of the *Populus* genome that contain genes controlling lignin content in shoots and roots and demonstrate that this combinational approach can be used to filter a candidate gene list to a substantially smaller subset of genes within a fixed confidence interval.

Materials and Methods

Description of QTLs

An F2 inbred interspecific hybrid poplar family was used to create a comprehensive genetic map containing 848 markers based on 293 segregating progeny as described by Yin et al. [40]. The overall observed genetic length was 1,927.6 cM. Phenotypic data was collected for all progeny using pyMBMS to obtain estimates of root lignin, root S/G ratio, stem lignin content, and stem S/G ratio [11, 32, 33]. MapQTL 5.0 was used to detect the underlying QTLs [38]. The establishment of genetic map, phenotying of lignin content, and S/G ratio of mapping individuals have been described by Yin et al. (in review).

Assigning Physical Position to the SSR Markers

Populus genome sequence, gene models, and functional categories for genes were downloaded from the JGI Populus Genome portal (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). The SSR primer resource, available at the International Poplar Genome Consortium website (http://www.ornl.gov/sci/ipgc/ssr_resource.htm; [34, 41]), was used to obtain the sequence information for the SSR primers and predicted SSR length. The physical position of each interval in the *Populus* genome was assigned based on BLAST results of the SSR primer nucleotide sequence against the genomic sequence. Additionally, the number of base pairs between the start of the left primer and end of right primer (according to BLAST result) had to be equivalent to the predicted length of the SSR marker. Perl script was used to automate this process. In total, 210 markers were successfully assigned physical position in the genome.

Assigning Physical Position to the QTLs

Assigning QTLs to physical positions in the genome was challenging as linear relationship between the physical and

genetic maps vary by position within the genome due to non-homogeneous distribution of chiasma across the genome. Thus, SSR markers flanking the QTLs were initially identified based on genetic positions. The relationship between genetic and physical distance for each QTL was then obtained by the ratio of physical distance between the markers and the genetic distance between them. This relationship was used to obtain the physical coordinates of the QTL in the genome by subtracting the difference between the ends of the QTL and the flanking marker.

Identification of Duplicated Genes Corresponding to the QTL Interval

Around 8,000 pairs of paralogous genes of similar age (excluding tandem duplications) were identified in the *Populus* genome. All genes in each QTL interval were identified based on the position of genes on each chromosome/linkage group (LG). The duplicated interval and corresponding duplicated gene information were then identified. Next, percent identity between a gene and its paralog was calculated using BLAST to align each pair. Finally, the best match *Arabidopsis* genes were identified by reciprocal blasting BLAST of the *Arabidopsis* gene set (TAIR Version 9) and *Populus* gene set to identify the top pair in each case.

Data Mining of Microarray Expression Profiles of Genes and Duplicated Genes

Populus balsamifera Affymetrix microarray datasets containing developmental tissue series (GSE13990 series) in GEO database at NCBI were used to examine the transcriptome level attributes of roots and differentiating xylem. We used this dataset to identify differences in gene expression between root and stem. The 50,848 probe sets with genome match correspond to 40,236 unique JGI *Populus trichocarpa* gene models. Cross-hybridizing and redundant probes for gene models as well as probes for alternatively spliced version of genes were eliminated in the analysis.

Identification of Differentially Expressed Genes

We used the RankProd package [16] to analyze the expression array data to identify differentially expressed genes. RankProd utilizes a rank product non-parametric method [6] to identify up- or down-regulated genes under differential conditions, e.g., two treatments, two tissue types, etc. The false discovery rate (FDR) value obtained was based on 10,000 random permutations [16]. The genes that had FDR values less than or equal to 0.10 were considered as differentially expressed.

Results

QTL Intervals

The QTL intervals for lignin and S/G ratio are located on seven linkage groups in the *Populus* genome (Fig. 1). The genetic position of each QTL interval is shown in Table 1. The QTL intervals for root lignin content were observed on LG II, LG VI, LG X, and LG XIV; for stem lignin content on LG II, LG VI, and LG X; for root S/G ratio on LG III, LG VI, LG X, and LG XIII; and for stem S/G ratio on LG II, LG III, LG VI, LG XIV, and LG VIII. These QTL intervals generally do not overlap, except on LG VI where QTL intervals for root and stem lignin content and root and stem S/G ratio co-localize and on LG X where QTL intervals for root and stem lignin content co-localize (Fig.1). The length of the QTL intervals ranged from 0.4 to 11 Mbp (Table 2); the majority of the QTL intervals were less than 2 Mbp in length. Correspondingly, the number of genes in the QTL intervals ranged from 44 to 1,501. The total number of genes in all the intervals was 4,530 (Supplementary Table 1). As there were two regions of overlapping QTLs, some genes were common to those QTLs, and the number of unique genes from all the QTL intervals was 3,788.

Duplication in *Populus* Genome and Duplicated Regions in QTL Interval

The *Populus* genome has undergone a recent genome-wide duplication event that has resulted in a conserved linear order of most of the genes within the duplicated chromosomal segments. QTLs on LG II have duplicated intervals on LG V; QTLs on LG III have duplicated intervals on LG V and scaffold_29; QTLs on LG VI have duplicated intervals on LG XVI and LG XVIII; QTLs on LG X have duplicated intervals on LG VIII (Fig. 2). Some intervals had higher numbers of genes conserved in the duplicated region as compared to the others. Across all intervals, on average, more than 53% of genes had retained a paralog in the duplicated interval and ranged from 25% to 80% (Table 3).

Comparison of Expression of Genes that Lie in the QTL Interval and Their Paralogs

Based on microarray evidence, 13 out of 19 QTL intervals were tissue specific, i.e., the QTL intervals corresponding to lignin content or S/G ratio were unique for either root or stem. Four of these QTL intervals, root lignin content (RL-1), stem lignin content (SL-3), root S/G ratio (RSG-1), and stem S/G ratio (SSG-5), were selected for a detailed analysis

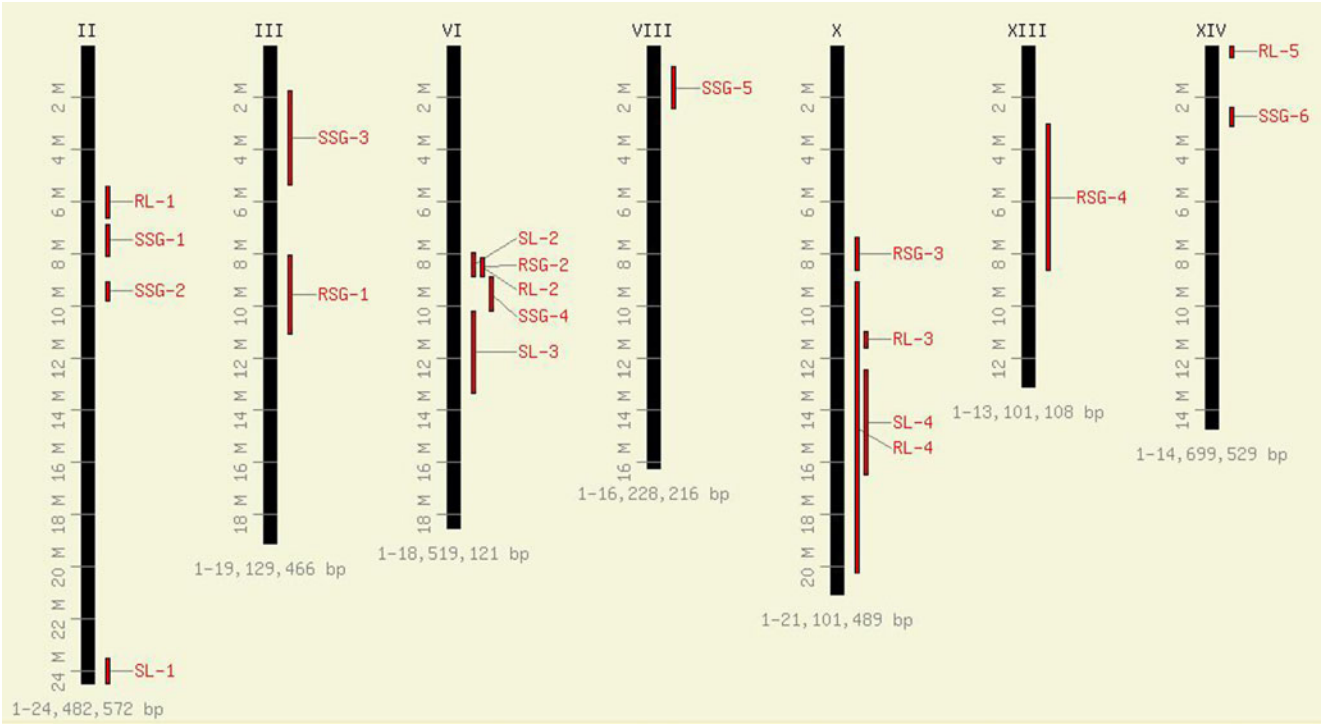

Fig. 1 Nineteen QTL interval intervals distributed on seven linkage groups (LG II, LG III, LG VI, LG VIII, LG X, LG XIII, and LG XIV) in *Populus*. *RL* root lignin QTL intervals, *RSG* root S/G ratio, *SL* stem lignin content, *SSG* stem S/G ratio

Table 1 Location of QTL intervals based on genetic map

QTL short name	QTL full name	LOD	% explanation	Linkage group	Peak position (cM)	1 LOD confidence (left, cM)	1 LOD confidence (right, cM)
RL-1	Root lignin content-1	2.77	18.2	II	62.13	55.13	67.785
RL-2	Root lignin content-2	5.06	8.9	VI	66.68	61.067	71.158
RL-3	Root lignin content-3	2.91	5.2	X	67.72	63.867	70.043
RL-4	Root lignin content-4	3.81	11.1	X	130.17	90.634	130.174
RL-5	Root lignin content-5	2.71	8.4	XIV	5.66	0	9.353
RSG-1	Root S/G ratio-1	5.19	14.9	III	57.82	48.823	65.823
RSG-2	Root S/G ratio-2	2.82	5.4	VI	71.16	62.624	71.158
RSG-3	Root S/G ratio-3	3.19	6	X	48.51	47.244	52.508
RSG-4	Root S/G ratio-4	3.21	6.6	XIII	50.39	39.936	60.298
SL-1	Stem lignin content-1	2.85	6.9	II	175.44	167.589	175.44
SL-2	Stem lignin content-2	4.29	7.6	VI	71.16	60.067	71.108
SL-3	Stem lignin content-3	4.21	7.3	VI	82.47	78.8	105.578
SL-4	Stem lignin content-4	2.71	4.6	X	80.15	75.202	105.634
SSG-1	Stem S/G ratio-1	5.02	14.7	II	76.8	71.759	82.132
SSG-2	Stem S/G ratio-2	5.51	12.5	II	90.76	88.759	95.09
SSG-3	Stem S/G ratio-3	3.22	6.4	III	23.62	10.724	32.197
SSG-4	Stem S/G ratio-4	2.58	5.2	VI	80.74	71.158	85.787
SSG-5	Stem S/G ratio-5	2.69	5.2	VIII	25.74	10.097	27.222
SSG-6	Stem S/G ratio-6	3.31	6.7	XIV	50.37	40.764	57.363

Table 2 Details of QTL in terms of physical distances

QTL short name	QTL full name	Linkage group	Final left flank physical position (bp)	Final right flank physical position (bp)	Length of interval (Mbp)
RL-1	Root lignin content-1	II	5,493,283	6,716,085	1.22
RL-2	Root lignin content-2	VI	8,063,866	8,911,035	0.85
RL-3	Root lignin content-3	X	10,873,565	11,681,744	0.81
RL-4	Root lignin content-4	X	9,141,424	20,293,167	11.15
RL-5	Root lignin content-5	XIV	1	485,878	0.49
RSG-1	Root S/G ratio-1	III	8,099,415	11,113,515	3.01
RSG-2	Root S/G ratio-2	VI	8,194,580	8,911,035	0.72
RSG-3	Root S/G ratio-3	X	7,291,775	8,702,806	1.41
RSG-4	Root S/G ratio-4	XIII	3,005,790	8,702,467	5.7
SL-1	Stem lignin content-1	II	23,535,453	24,482,567	0.95
SL-2	Stem lignin content-2	VI	7,979,913	8,906,838	0.93
SL-3	Stem lignin content-3	VI	10,229,422	13,356,289	3.13
SL-4	Stem lignin content-4	X	12,428,930	16,367,409	3.94
SSG-1	Stem S/G ratio-1	II	6,929,755	7,993,901	1.06
SSG-2	Stem S/G ratio-2	II	9,110,752	9,878,601	0.77
SSG-3	Stem S/G ratio-3	III	1,810,361	5,435,304	3.62
SSG-4	Stem S/G ratio-4	VI	8,911,035	10,139,184	1.23
SSG-5	Stem S/G ratio-5	VIII	875,268	2,359,766	1.48
SSG-6	Stem S/G ratio-6	XIV	2,282,351	3,052,827	0.77

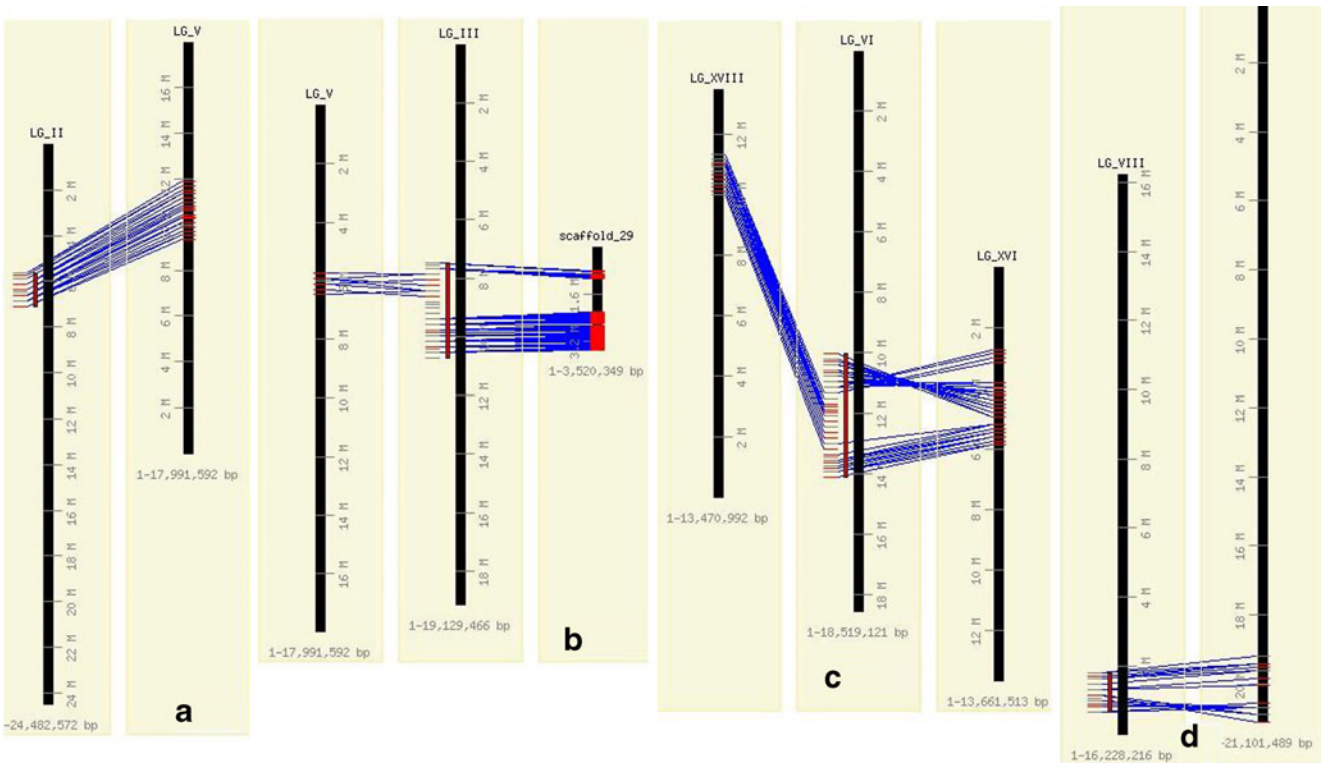

Fig. 2 QTL interval and display of duplication in genes in the interval for **a** root lignin content-1, **b** root S/G ratio-1, **c** stem lignin content-3, and **d** stem S/G ratio. Each *blue line* represents a gene and its paralog in the duplicated region

Table 3 Details of duplication and differences in expression

QTL	Total number of genes	Has paralogs	% id>90	% id>90 and expression in tissue of QTL	% id<90	% id<90 and expression in tissue of QTL	Missing paralogs and expression in tissue of QTL
Number of genes in each category							
RSG-3	118	51	16	2	35	5	13
SL-3	247	171	65	3	106	6	8
SL-2	61	14	7		7		3
SSG-3	222	136	50	3	86	3	6
RSG-4	454	209	80	6	129	14	33
SL-1	88	25	7		18		7
SSG-6	115	69	26	3	43	4	2
RSG-1	278	170	63	1	107	4	8
RL-3	98	71	30	2	41	5	5
SSG-5	226	155	67	7	88	10	6
SSG-4	66	29	10		19		3
SL-4	548	332	119	11	213	14	16
RL-4	1,501	981	391	33	590	62	74
RL-2	52	14	7		7		12
RL-5	54	27	10	2	17	3	4
RL-1	138	94	39	4	55	5	2
SSG-1	123	51	20	3	31		8
SSG-2	97	77	24	2	53	2	
RSG-2	44	14	7		7		11

because they occurred in paralogous regions and contained differential tissue data from microarray experiments.

In order to use the above data to filter the candidate gene list within each of the selected QTL intervals, three alternative approaches were used to integrate duplication information and differential expression of paralogs (Fig. 3). First, filtering was based on non-duplicated genes within a QTL and those which have higher expression in tissue related to the QTL. Second, differential microarray results were used to identify genes within the interval with expression evidence in the identified tissue whose paralogous genes did not display expression in the corresponding tissue. For example, we identified genes, present in QTL interval for stem lignin content, that show higher expression in xylem relative to root as compared to gene expression of paralogs. In addition, the genes within the QTL interval that have a predicted role in cell wall biosynthesis (e.g., PAL, 4CL, etc.) were promoted to the candidate gene list.

Genes in the QTL Interval

Root Lignin Content The total number of genes in the RL-1 interval on LG II was 138. Out of these, 94 had paralogs and 44 genes did not retain paralogs in the duplicated interval (Table 3). Two of these non-duplicated genes with higher expression in root were calmodulin (eugene3.00020820) and a signal transduction response regulator gene (gw1.II.42.1; Table 4). Five duplicated genes with <90% similarity and higher expression in root were replication factor (estExt_fgenesh4_pg.C_LG_II0728), an auxin response factor (fgenesh4_pg.C_LG_II000830), a glycosyl hydrolase hydrolyzing *o*-glycosyl compound (fgenesh4_pg.C_LG_II000867), a zinc finger transcription factor (grail3.0003072701), and an exoribonuclease (gw1.II.1849.1; Table 5). Four duplicated genes with >90% similarity and higher expression in root were sulfate transporter (eugene3.00020855), alcohol dehydrogenase (fgenesh4_pg.C_LG_II000742), a NAC domain protein (grail3.0003068301), and a nodulin-like protein (gw1.II.1386.1; Table 6). In total, the number of genes within the interval was filtered down from 138 equally likely candidates to 11 with supportive duplication and/or expression evidence.

Stem Lignin Content The total number of genes in the SL-3 interval on LG VI was 247. Out of these, 171 had paralogs and 76 genes did not retain paralogs in the duplicated interval (Table 3). Eight of these non-duplicated genes with higher expression in xylem were hypothetical protein (estExt_fgenesh4_pm.C_LG_VI0468), proteins with no known function and unique to Populus (eugene3.00061181, fgenesh4_pg.C_LG_VI001243), a nucleic acid binding

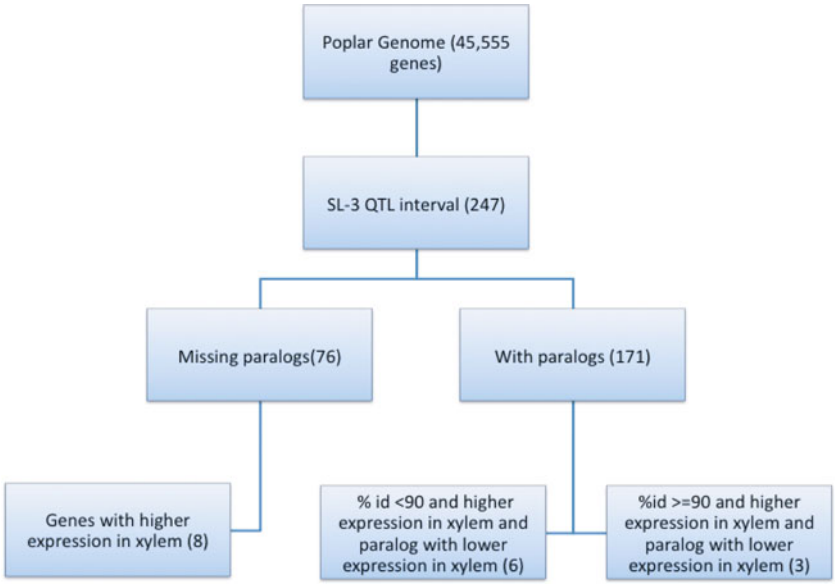

Fig. 3 Process of filtering genes. *SL-3* stem lignin QTL interval

(eugene3.00061373), protein associated with CCR4 transcription complex (grail3.0030015901), a plastocyanin-like domain-containing protein (gw1.VI.2580.1), a ribosomal protein (gw1.VI.2649.1), and a peptidyl-prolyl *cis–trans* isomerase, cyclophilin type protein (gw1.VI.847.1; Table 4). Six duplicated genes with <90% similarity and higher expression in xylem were kinesin protein involved in microtubule-based movement (estExt_fgenesh4_pm.

Table 4 List of genes that did not have a paralogous gene model in duplicated region and that showed higher expression in the tissue related to the QTL

QTL	Gene	% id with paralog	Function	FDR
RL-1	eugene3.00020820	NA	Calmodulin	≤0.001
RL-1	gw1.II.42.1	NA	Two-component response regulator	≤0.1
RSG-1	estExt_fgenesh4_pg.C_LG_III0677	NA	Oligopeptide transporter	≤0.01
RSG-1	estExt_Genewise1_v1.C_LG_III1770	NA	RNA helicase	≤0.1
RSG-1	eugene3.00030584	NA	Peroxidase	≤0.1
RSG-1	eugene3.00030600	NA	Formin-like protein	≤0.1
RSG-1	fgenesh4_pg.C_LG_III000886	NA	Expressed protein	≤0.1
RSG-1	grail3.0018015801	NA	ATP-dependent RNA helicase	≤0.01
RSG-1	gw1.III.1044.1	NA	DNA J protein	≤0.1
RSG-1	gw1.III.2608.1	NA	Transporter-like protein	≤0.001
SL-3	estExt_fgenesh4_pm.C_LG_VI0468	NA	Hypothetical protein	≤0.01
SL-3	eugene3.00061181	NA	No hits	≤0.1
SL-3	eugene3.00061373	NA	Nucleic acid-binding	≤0.1
SL-3	fgenesh4_pg.C_LG_VI001243	NA	No hits	≤0.1
SL-3	grail3.0030015901	NA	Associated with CCR4 transcription complex	≤0.01
SL-3	gw1.VI.2580.1	NA	Plastocyanin-like domain-containing protein	≤0.01
SL-3	gw1.VI.2649.1	NA	Ribosomal protein	≤0.1
SL-3	gw1.VI.847.1	NA	Peptidyl-prolyl *cis–trans* isomerase	≤0.01
SSG-5	eugene3.00080178	NA	Unknown protein	≤0.01
SSG-5	eugene3.00080195	NA	No hits	≤0.1
SSG-5	eugene3.00080203	NA	Ankyrin repeat family, transmembrane transport	≤0.001
SSG-5	eugene3.00080273	NA	No hits	≤0.001
SSG-5	fgenesh4_pg.C_LG_VIII000212	NA	Cytochrome P450	≤0.001
SSG-5	gw1.VIII.950.1	NA	Vacuolar protein	≤0.1

Table 5 List of genes that have a paralogous gene model; and the % identity is less than 90%, and that showed higher expression in the tissue related to the QTL

QTL	Gene	% id with paralog	Function	FDR
RL-1	estExt_fgenesh4_pg.C_LG_II0728	59	Replication factor	≤0.1
RL-1	fgenesh4_pg.C_LG_II000830	84	Auxin response factor	≤0.1
RL-1	fgenesh4_pg.C_LG_II000867	75	Glycosyl hydrolase (xylosidase)	≤0.1
RL-1	grail3.0003072701	69	Zinc finger transcription factor	≤0.01
RL-1	gw1.II.1849.1	89	Exoribonuclease	≤0.1
RSG-1	fgenesh4_pg.C_LG_III000669	87	calcium transporting ATPase	≤0.01
RSG-1	grail3.0018007101	67	Glucosyl transferase	≤0.001
RSG-1	grail3.0018018701	80	Tetratricopeptide-containing protein	≤0.01
RSG-1	gw1.III.1613.1	78	Proline-rich protein	≤0.01
SL-3	estExt_fgenesh4_pm.C_LG_VI0481	88	Kinesin (microtubule-based movement)	≤0.01
SL-3	estExt_fgenesh4_pm.C_LG_VI0500	51	Hypothetical protein	≤0.001
SL-3	estExt_Genewise1_v1.C_LG_VI2154	44	Senescence-associated protein	≤0.1
SL-3	grail3.0030003201	73	Unknown protein	≤0.01
SL-3	grail3.0030006902	81	Unknown protein	≤0.001
SL-3	gw1.VI.781.1	85	Nodulin-like protein	≤0.001
SSG-5	eugene3.00080177	75	Germin-like protein	≤0.1
SSG-5	eugene3.00080251	68	Acyl-CoA-binding family protein	≤0.1
SSG-5	eugene3.00080330	81	GATA zinc fringe protein	≤0.001
SSG-5	fgenesh4_pg.C_LG_VIII000250	67	Unknown protein	≤0.01
SSG-5	fgenesh4_pg.C_LG_VIII000264	77	Unknown protein	≤0.1
SSG-5	fgenesh4_pm.C_LG_VIII000069	81	Unknown protein	≤0.001
SSG-5	fgenesh4_pm.C_LG_VIII000111	81	Pumilio-family RNA-binding protein	≤0.01
SSG-5	grail3.0049006403	88	Chorismate synthase	≤0.1
SSG-5	gw1.VIII.1321.1	83	Pectate lyase-like protein	≤0.001
SSG-5	gw1.VIII.1497.1	85	Acyl-CoA-binding family protein	≤0.1

Table 6 List of genes that have a paralogous gene model; and the % identity is greater than 90%, and that showed higher expression in the tissue related to the QTL

QTL	Gene	% id with paralog	Function	FDR
RL-1	eugene3.00020855	90	Sulfate transporter	≤0.01
RL-1	fgenesh4_pg.C_LG_II000742	94	Alcohol dehydrogenase	≤0.001
RL-1	grail3.0003068301	91	NAC domain protein	≤0.001
RL-1	gw1.II.1386.1	93	Nodulin-like protein	≤0.001
RSG-1	fgenesh4_pg.C_LG_III000900	93	WRKY family transcription factor	≤0.001
SL-3	estExt_fgenesh4_pg.C_LG_VI1102	90	Endonuclease/exonuclease hydrolase activity	≤0.01
SL-3	eugene3.00061209	96	Expressed protein	≤0.01
SL-3	eugene3.00061339	93	UDP-D-Glucuronate 4-epimerase, nucleotide sugar interconversion pathway	≤0.1
SSG-5	estExt_fgenesh4_pg.C_LG_VIII0179	90	expressed protein	≤0.01
SSG-5	estExt_fgenesh4_pm.C_LG_VIII0087	91	Glucosyl transferase, cellulose synthase-like	≤0.001
SSG-5	eugene3.00080299	92	Vacuolar protein, vacuolar biogenesis	≤0.1
SSG-5	eugene3.00080329	90	Pleckstrin homology (PH) domain-containing protein	≤0.01
SSG-5	grail3.0049010802	97	CCAAT-box-binding transcription factor	≤0.1
SSG-5	gw1.VIII.1083.1	91	Photoreceptor-interacting protein	≤0.001
SSG-5	gw1.VIII.2327.1	93	Exostosin family, GT 47	≤0.01

C_LG_VI0481), a hypothetical protein (estExt_fgenesh4_pm.C_LG_VI0500), a senescence associated protein (estExt_Genewise1_v1.C_LG_VI2154) and unknown proteins (grail3.0030003201, grail3.0030006902), and a nodulin-like protein (gw1.VI.781.1; Table 5). Three duplicated genes with >90% similarity and higher expression in xylem were protein with hydrolase activity (estExt_fgenesh4_pg.C_LG_VI1102), an expressed protein with no known function (eugene3.00061209), and a UDP-D-glucuronate 4-epimerase involved in nucleotide sugar interconversion pathway (eugene3.00061339; Table 6). In total, the number of genes within the interval was filtered down from 247 equally likely candidates to 17 with supportive duplication and/or expression evidence.

Root S/G ratio The total number of genes in the RSG-1 interval on LG III was 278. Out of these, 170 had paralogs and 108 genes did not retain paralogs in the duplicated interval (Table 3). Eight of these non-duplicated genes with higher expression in root were oligopeptide transporter (estExt_fgenesh4_pg.C_LG_III0677), a RNA helicase protein (estExt_Genewise1_v1.C_LG_III1770, grail3.0018015801), a peroxidase (eugene3.00030584), a formin-like protein (eugene3.00030600), an expressed protein (fgenesh4_pg.C_LG_III000886), a DNA J protein (gw1.III.1044.1), and a transporter-like protein (gw1.III.2608.1; Table 4). Four duplicated genes with <90% similarity and higher expression in root were calcium transporting ATPase (fgenesh4_pg.C_LG_III000669), a glucosyl transferase (grail3.0018007101), a tetratricopeptide-containing protein (grail3.0018018701), and a proline-rich protein (gw1.III.1613.1; Table 5). One duplicated gene with >90% similarity and higher expression in root was WRKY family transcription factor (fgenesh4_pg.C_LG_III000900; Table 6). In total, the number of genes within the interval was filtered down from 278 equally likely candidates to 13 with supportive duplication and/or expression evidence.

Stem S/G ratio The total number of genes in the SSG-5 interval on LG VIII was 226. Out of these, 155 had paralogs and 71 genes did not retain paralogs in the duplicated interval (Table 3). 2Six of these non-duplicated genes with higher expression in xylem were unknown protein (eugene3.00080178), a protein unique to *Populus* (eugene3.00080195, eugene3.00080273), an ankyrin repeat family involved in transmembrane transport (eugene3.00080203), a cytochrome P450 protein (fgenesh4_pg.C_LG_VIII000212), and a vacuolar protein (gw1.VIII.950.1; Table 4). Ten duplicated genes with <90% similarity and higher expression in xylem were germin-like protein (eugene3.00080177), an acyl-CoA-binding family protein (eugene3.00080251), a GATA zinc finger protein (eugene3.00080330), unknown proteins (fgenesh4_pg.C_LG_VIII000250, fgenesh4_pg.C_LG_VIII000264, fgenesh4_pm.C_LG_VIII000069), a pumilio-family RNA-binding protein (fgenesh4_pm.C_LG_VIII000111), a chorismate synthase (grail3.0049006403), a pectate lyase-like protein (gw1.VIII.1321.1), and an acyl-CoA-binding family protein (gw1.VIII.1497.1; Table 5). Seven duplicated gene with >90% similarity and higher expression in xylem were expressed protein (estExt_fgenesh4_pg.C_LG_VIII0179), a glucosyl transferase also annotated as cellulose synthase-like (estExt_fgenesh4_pm.C_LG_VIII0087), a vacuolar protein (eugene3.00080299), pleckstrin homology domain-containing protein (eugene3.00080329), a CCAAT-box-binding transcription factor (grail3.0049010802), a photoreceptor-interacting protein (gw1.VIII.1083.1), and an exostosin family protein also annotated as glucoside transferase 47 (gw1.VIII.2327.1; Table 6). In total, the number of genes within the interval was filtered down from 226 equally likely candidates to 23 with supportive duplication and/or expression evidence.

Discussion

Whole-genome duplication events, followed by extensive genome reorganization, chromosomal rearrangements, and gene loss, have been widespread during the evolution of plants [31]. As a consequence of duplication, paralogs created in the genome may have one of several possible fates, including non-functionalization, neo-functionalization, and sub-functionalization [18]. The most acknowledged mode is non-functionalization where one of the copies loses function or is silenced, resulting in a pseudogene [1]. The process of neo-functionalization, where one ancestral copy retains its function and the other is free to accumulate mutations, results in acquisition of novel function. During the process of sub-functionalization the sister copies, i.e. paralogs, show different but overlapping functions [18]. Here, some duplicated genes show differential expression among organs within a single plant. In recent allopolyploidization in cotton some genes were silenced in one organ with respect to another. Similar outcomes were detected in artificial allopolyploidization [2]. In *Arabidopsis* there is evidence of sub-functionalization where clusters of duplicated genes show evidence of concerted divergence in their expression in an organ-specific expression [4].

The *Populus* genome has undergone multiple genome-wide duplications [35]. The Salicoid duplication currently contains around 8,000 pairs of genes that are syntenous across mega-base regions of the genome. As a result, almost every segment in the *Populus* genome has a parallel paralogous interval somewhere else in the genome. Yet,

QTL intervals for many stem and root lignin and S/G ratio phenotypes are present in only one position (Fig. 1). This suggests that different sets of genes to root and stem QTLs providing an opportunity to leverage the segmental duplication information. Along with gene expression of paralogous gene intervals, higher likelihood values can be assigned to genes or gene sets that are functionally related to the measured phenotype.

This expansive gene declaration results in each QTL interval having hundreds to thousands of genes. Our filtering approach led to a reduced set of genes, most not previously reported to play a direct role in monolignol biosynthesis. These filtered genes included regulatory proteins that may have roles in cell wall formation, vascular transport, and unknown function. For example, in the root lignin interval, a NAC domain transcription factor (grail3.0003068301) is present, and NAC domain transcription factors have been implicated as key regulator of secondary cell wall synthesis in *Arabidopsis* [43]. A signal transduction response regulator (gw1.II.42.1) was also identified in this interval and is an ideal candidate for further transgenic work as is kinesin (estExt_fgenesh4_pm.C_LG_VI0481), which is involved in the oriented deposition of cellulose microfibrils in *Arabidopsis* [42]. In the root S/G ratio interval two proteins seem very promising. One is glycosyl hydrolase (fgenesh4_pg.C_LG_II000867) and the other is UDP-D-glucuronate 4-epimerase (eugene3.00061339) involved in nucleotide sugar interconversion pathways. In the stem S/G content interval a cytochrome P450 (fgenesh4_pg.C_LG_ VIII000212) is a good candidate for further experimental work. Exostosin gene (gw1.VIII.2327.1) has also been shown to be more highly co-regulated with cellulose synthase genes in *Arabidopsis* [21].

The filtered gene set provides a feasible opportunity to determine gene function via functional genetics work. That is, based on the computational approach described above, a set of 15–20 candidate genes can be used in RNAi knockdown experiments, mutant complementation experiments, and in association genetics studies correlating single nucleotide polymorphisms frequency and measured phenotypes. An integrated approach that combines QTL mapping with fine-scale mapping using association mapping would require investigating SNPs associated with trait using SNP arrays [17]. Multiple SNPs need to be assayed per gene as linkage disequilibrium decays rapidly in *Populus*. The filtration strategy discussed in this paper can be used to select candidate genes to assay SNPs and uncover the underlying DNA polymorphism associated with lignin content and lignin S/G ratio.

The unique approach of filtering for genes based on duplication evidence and expression data of paralogs has its limitations. The assembly of the *Populus* genome is still in a draft state and has numerous captured gaps where the length of the missing fragment is known and non-captured gaps where the length of the missing fragment is not known. Moreover, the lack of microarray datasets for *Populus* compared to *Arabidopsis* is also a limiting factor. As more microarray datasets become available, more robust statistical analyses will be feasible. The design of Affymetrix microarray adds to the challenge. Due to the overlapping nature of the Affymetrix probe sets it is frequently difficult to distinguish paralogs. Due to the lack of the microarray datasets, we based our analysis on microarray datasets from *P. balsamifera* on Affymetrix chips, whereas the QTL intervals were obtained from *P. trichocarpa*. Future studies of this nature should use the expression data from individuals with extreme phenotypes in the population used to detect QTL.

Conclusions

This paper provides a computational approach for integrating QTLs with expression data and *Populus* genome duplication information to assign higher likelihood values to candidate genes with greater precision than other. The analytical approach was successful in identifying both genes of suspected cell wall biosynthetic function as well as genes of putative cell wall biosynthetic function. Genes of unknown or putative functions would most likely not have been examined without such an approach. In total, the list of genes in QTL intervals was reduced from hundreds or thousands of genes to 15–20 genes. These results provide a roadmap for future experimental work attempting to discover cell wall recalcitrance genes and the ultimate utility of plant biomass as an energy feedstock.

Acknowledgments The authors would like to thank Dr. Stan Wullschleger, David Weston, Lee Gunter, and Manojit Basu for their technical reviews of this paper. The present study was enabled by research funds through the BioEnergy Science Center, which is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the US Department of Energy

References

1. Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8:135–141
2. Adams KL, Percifield R, Wendel JF (2004) Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics 168:2217–2226

3. Bechtold N, Ellis J, Pelletier G (1993) In planta Agrobacterium-mediated gene transfer by infiltration of adult Arabidopsis thaliana plants. C R Acad Sci Ser III Sci Vie 316:1194–1199
4. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679–1691
5. Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annu Rev Plant Biol 54:519–546
6. Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573:83–92
7. Brodeur-Campbell SE, Vucetich JA, Richter DL, Waite TA, Rosemier JN, Tsai CJ (2006) Insect herbivory on low-lignin transgenic aspen. Environ Entomol 35:1696–1701
8. Byrne KP, Wolfe KH (2007) Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. Genetics 175:1341–1350
9. Chen F, Dixon RA (2007) Lignin modification improves fermentable sugar yields for biofuel production. Nat Biotechnol 25:759–761
10. Chiang VL (2006) Monolignol biosynthesis and genetic engineering of lignin in trees, a review. Environmental Chemistry Letters 4 (3):143–146
11. Davis MF, Tuskan GA, Payne P, Tschaplinski TJ, Meilan R (2006) Assessment of *Populus* wood chemistry following the introduction of a Bt toxin gene. Tree Physiology 26:557–564
12. Davison BH, Drescher SR, Tuskan GA, Davis MF, Nghiem NP (2006) Variation of S/G ratio and lignin content in a *Populus* family influences the release of xylose by dilute acid hydrolysis. Appl Biochem Biotechnol 130(1–3):427–435
13. Dinus RJ, Payne P, Sewell NM, Chiang VL, Tuskan GA (2001) Genetic modification of short rotation poplar wood: properties for ethanol fuel and fiber productions. Crit Rev Plant Sci 20:51–69
14. Farrokhi N, Burton RA, Brownfield L, Hrmova M, Wilson SM, Bacic A, Fincher GB (2006) Plant cell wall biosynthesis: genetic, biochemical and functional genomics approaches to the identification of key genes. Plant Biotechnol J 4:145–167
15. Himmel ME, Ding SY, Johnson DK, Adney WS, Nimlos MR, Brady JW et al (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 315(5813):804–807
16. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 22:2825–2827
17. Ingvarsson PK, Garcia V, Luquez V, Hall D, Jansson S (2008) Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, Salicaceae). Genetics 178:2217–2226
18. Jaillon O, Aury JM, Wincker P (2009) "Changing by doubling", the impact of whole genome duplications in the evolution of eukaryotes. Comptes Rendus Biologies 332:241–253
19. Kalluri UC, Joshi CP (2004) Differential expression patterns of two cellulose synthase genes are associated with primary and secondary cell wall development in aspen trees. Planta 220:47–55
20. Leple JC, Dauwe R, Morreel K, Storme V, Lapierre C, Pollet B et al (2007) Downregulation of cinnamoyl-coenzyme A reductase in poplar: multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. Plant Cell 19:3669–3691
21. Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci USA 102:8633–8638
22. Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA et al (2006) The path forward for biofuels and biomaterials. Science 311(5760):484–489
23. Ralph J, Akiyama T, Kim H, Lu FC, Schatz PF, Marita JM et al (2006) Effects of coumarate 3-hydroxylase down-regulation on lignin structure. J Biol Chem 281:8843–8853
24. Ranjan P, Kao YY, Jiang H, Joshi CP, Harding SA, Tsai CJ (2004) Suppression subtractive hybridization-mediated transcriptome analysis from multiple tissues of aspen (*Populus tremuloides*) altered in phenylpropanoid metabolism. Planta 219:694–704
25. Rikke BA, Johnson TE (1998) Towards the cloning of genes underlying murine QTLs. Mamm Genome 9:963–968
26. Roberts AW, Bushoven JT (2007) The cellulose synthase (CESA) gene superfamily of the moss *Physcomitrella patens*. Plant Mol Biol 63:207–219
27. Ron M, Weller JI (2007) From QTL to QTN identification in livestock—winning by points rather than knock-out: a review. Anim Genet 38:429–439
28. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D et al (2007) Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 308B:58–73
29. Rubin EM (2008) Genomics of cellulosic biofuels. Nature 454 (7206):841–845
30. Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. Trends Plant Sci 10:297–304
31. Semon M, Wolfe KH (2007) Consequences of genome duplication. Curr Opin Genet Dev 17:505–512
32. Sykes R, Kodrzycki B, Tuskan G, Foutz K, Davis M (2008) Within tree variability of lignin composition in *Populus*. Wood Sci Technol 42:649–661
33. Tuskan GA, West D, Bradshaw HD, Neale D, Sewell M, Wheeler N et al (1999) Two high-throughput techniques for determining wood properties as part of a molecular genetics analysis of loblolly pine and hybrid poplar. Appl Biochem Biotech 77–79:1–11
34. Tuskan GA, Gunter LE, Yang ZMK, Yin TM, Sewell MM, DiFazio SP (2004) Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa*. Can J For Res 34:85–93
35. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604
36. Van de Peer Y (2004) Computational approaches to unveiling ancient genome duplications. Nat Rev, Genet 5:752–763
37. Vanholme R, Morreel K, Ralph J, Boerjan W (2008) Lignin engineering. Curr Opin Plant Biol 11:278–285
38. Van Ooijen JW (2004) MapQTL 5, software for the mapping of quantitative trait loci in experimental populations. Kyazma B.V, Wageningen
39. Yang XH, Jawdy S, Tschaplinski TJ, Tuskan GA (2009) Genome-wide identification of lineage-specific genes in *Arabidopsis*, *Oryza* and *Populus*. Genomics 93:473–480
40. Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA (2004) Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. Theor Appl Genet 109:451–463
41. Yin TM, Zhang XY, Gunter LE, Li SX, Wullschleger SD, Huang MR et al (2009) Microsatellite primer resource for Populus developed from the mapped sequence scaffolds of the Nisqually-1 genome. New Phytol 181(2):498–503
42. Zhong R, Burk DH, Morrison WH, Ye ZH (2002) A kinesin-like protein is essential for oriented deposition of cellulose microfibrils and cell wall strength. Plant Cell 14:3101–3117
43. Zhong R, Demura T, Ye ZH (2006) SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. Plant Cell 18:3158–3170