Exome resequencing and GWAS for growth, ecophysiology, and chemical and metabolomic composition of wood of Populus trichocarpa

Fernando P. Guerra
Universidad de Talca https://orcid.org/0000-0001-7174-9738
Haktan Suren
Virginia Polytechnic Institute and State University
Jason Holliday
Virginia Polytechnic Institute and State University
James H. Richards
University of California Davis
Oliver Fiehn
University of California Davis
Randi Famula
University of California Davis
Brian J. Stanton
GreenWood Resources
Richard Shuren
GreenWood Resources
Robert Sykes
National Renewable Energy Laboratory
Mark F. Davis
National Renewable Energy Laboratory
David B. Neale (dbneale@ucdavis.edu)

Research article

Keywords: Populus; GWAS; sequence capture; growth; stable isotopes; lignin; cellulose; wood metabolome

Posted Date: October 12th, 2019

DOI: https://doi.org/10.21203/rs.2.9589/v2

Version of Record: A version of this preprint was published on November 20th, 2019. See the published version at https://doi.org/10.1186/s12864-019-6160-9.

Abstract

Background: Populus trichocarpa is an important forest tree species for the generation of lignocellulosic ethanol. Understanding the genomic basis of biomass production and chemical composition of wood is fundamental in supporting genetic improvement programs. Considerable variation has been observed in this species for complex traits related to growth, phenology, ecophysiology and wood chemistry. Those traits are influenced by both polygenic control and environmental effects, and their genome architecture and regulation are only partially understood. Genome wide association studies (GWAS) represent an approach to advance that aim using thousands of single nucleotide polymorphisms (SNPs). Genotyping using exome capture methodologies represent an efficient approach to identify specific functional regions of genomes underlying phenotypic variation. Results: We identified 813K SNPs, which were utilized for genotyping 461 P. trichocarpa clones, representing 101 provenances collected from Oregon and Washington, and established in California. A GWAS performed on 20 traits, considering single SNP-marker tests identified a variable number of significant SNPs (p-value < 6.1479E-8) in association with diameter, height, leaf carbon and nitrogen contents, and δ 15 N. The number of significant SNPs ranged from 2 to 220 per trait. Additionally, multiple-marker analyses by sliding-windows tests detected between 6 and 192 significant windows for the analyzed traits. The significant SNPs resided within genes that encode proteins belonging to different functional classes as such protein synthesis, energy/metabolism and DNA/RNA metabolism, among others. Conclusion: SNP-markers within genes associated with traits of importance for biomass production were detected. They contribute to characterize the genomic architecture of P. trichocarpa biomass required to support the development and application of marker breeding technologies.

Background

Populus species and their hybrids are suitable feedstocks for second-generation biofuel production due to their rapid growth rates and favorable cell wall chemistry [1, 2]. In particular, the model species *Populus trichocarpa* Torr. & A. Gray (black cottonwood), native to western North America, has been used in breeding for generating commercial cultivars [3]. Biomass yield and chemical quality of *P. trichocarpa* cultivars, as well as their improvement, depend on multiple biological and environmental factors [4]. Considerable phenotypic and genetic variation has been observed in *P. trichocarpa* for complex traits related to growth, phenology, morphology, ecophysiology and wood chemistry [5-10]. These phenotypes include diameter and height [11, 12], bud set and flush [6, 13, 14], leaf morphology [15], water-use efficiency (WUE) [16, 17], secondary xylem composition [18] and wood metabolome [5]. This sort of traits has been also correlated with environmental variables such as latitude, daylength and temperature [5, 6, 14-16, 19].

Association analyses based on SNPs have been applied in recent years to identify polymorphisms controlling variation in complex traits of interest for biofuel production in *Populus* species [9, 15, 18-21]. Different approaches (candidate gene or GWAS) as well as genotyping platforms have been used, with single SNP-markers accounting for, in general, a low percentage of the phenotypic variation (1-8 %) in

studied traits. These results support the polygenic nature and complexity of inheritance patterns and justifies increasing efforts to elucidate the genomic basis controlling those phenotypes.

Among “next-generation” sequencing alternatives, genome complexity reduction by sequence capture, or targeted sequencing, represents an efficient approach to performing genome wide analysis [22]. This method restricts attention only to specific genome regions (both genic and intergenic) of interest for molecular breeding as well as investigations into the diversity, population structure and demographic history of unstructured natural populations among others [23]. This approach has advantage of being quick, simple, and requires relatively small amount of input DNA [24]. Furthermore, compared with alternatives such as whole genome sequencing, it is reduced in terms of non-pertinent repetitive sequences, allows multiplexing of more samples for a given sequencing space, identifies functional molecular markers, provides high coverage for identification of low frequency sequence variants, and can circumvent problems arising from the presence of paralogous genes derived from duplication or polyploidization events [24]. This is particularly important for *Populus* species, which have experienced a whole-genome duplication event [25]. It was demonstrated by the application of an exome capture approach for analyzing the genomic architecture of clinal variation in *P. trichocarpa* [26].

In the present study, we employ sequence capture for genotyping and performing a GWAS in a *P. trichocarpa* population of 461 clones from 101 provenances collected from the Pacific Northwest (Oregon and Washington) in the United States. In an previous study [5], representatives of these clones were established in a clonal trial in California and characterized, both by traditional field measurements and high-throughput phenotyping, in describing a suite of traits involved in biomass production and wood chemical composition. Now, we coupled these phenotypic measures with specific exome capture-based genotyping to identify SNPs underlying observed trait variation. The association population was generated with germplasm collected from the southern part of the *P. trichocarpa* range in North America, and it was established and evaluated (at age two) in a trial located significantly to the south than that range. That represents a particular environmental/experimental condition, useful to determine, for example, the effects of geographic relocation on the *P. trichocarpa* performance. Understanding genetic variation at a genome-wide scale is fundamental for developing genome-based breeding technologies suitable for supporting the development of genetically improved plantations for bioethanol production.

Methods

Association population, growth conditions and phenotypic characterization

The association population was comprised of a set of 461 *P. trichocarpa* clones. These represented 101 provenances, within 14 river systems located west of the Cascade Mountains in Oregon and Washington between 48°54’ N latitude (Nooksack River, Whatcom County, Washington) and 43°47’ N latitude (Middle Fork, Willamette River, Lane County, Oregon) collected by GreenWood Resources [18]. A clonal trial was established at the University of California, Davis, California (38°32’42” N, 121°47’42” W) for phenotypic

measurements and sample collection. The experiment was located on a silt loam soil (Entisol, Yolo series). Plants were produced from rooted containerized cuttings and established in April 2009, in an array of 1.83 x 1.83 m, following a randomized block design, with three blocks and one ramet per clone per block. Blocks were considered to control a north–south variation in the trial, associated with non-homogenous soil conditions. From 2009 to 2011, plants were irrigated once a week from June to September each year. No fertilizer was applied during the study.

The clonal trial was already characterized by Guerra *et al.* [5], in terms of traits dealing with growth, spring bud phenology, ecophysiology, and the chemical composition and metabolome of wood. Table 1 summarizes main statistics and clonal repeatability (heritability) estimates for those traits. A brief description about their measurements is indicated in the following subsections. Individual broad-sense heritability (H^2_i) was estimated using formula (1) [5]: (See Formula 1 in the Supplementary Files)

where σ^2_g, σ^2_c, σ^2_{fxb} and σ^2_e represent the variance due the genetic cluster, clone within cluster, cluster x block and residual, respectively, and $\sigma^2_{ee'}$ is the covariance between residuals of the same clone in two blocks.

Data used in the phenotypic characterization is available at the TreeGenes platform (http://dendrome.ucdavis.edu/treegenes/) under the accession number TGDR050.

Diameter, height, volume and spring-bud phenology measurements

Three growth parameters (diameter at breast height [DBH], total height [h] and volume index [Vol]), at age two, were measured in October 2010, as described by Guerra *et al.* [5]. In particular, Vol was estimated as $Vol = \pi (DBH/2)^2 \times h$ [13]. Additionally, days to bud flush (DBF), were recorded every three days from March to April 2011, as indicated previously [5].

Chemical composition and metabolome of wood

Wood cores collected from tree stems (0.3 m above ground level), at age three, in September 2011, were utilized for analyzing the chemical composition and metabolome of wood [5]. The content of 5 and 6-carbon sugars and lignin, as well as the syringyl:guaiacyl monolignol ratio (S:G) were determined from wood cores by high-throughput pyrolysis molecular beam mass spectrometry (pyMBMS), at the National Renewable Energy Laboratory (Golden, CO, USA). Simultaneously, wood metabolites were quantified from another set of wood cores by gas chromatography coupled with time-of-flight mass spectrometry (GC-TOF-MS), at the West Coast Metabolomics Center, at UCDavis. Methodological details have been described elsewhere [5]. For association analysis, five metabolites were selected. These corresponded to those with the highest estimates of heritabilities according to Guerra *et al.* [5]: 4-hydroxybenzoic acid (HbA;), galactinol (Gal;), adenosine (Ade;), galactonic acid (GAc;), and alpha tocopherol (Toc;). These estimates involved a significant genetic variation across the analyzed genotypes, indicating their importance on wood composition variation. They were selected to maximize the power for detecting significant associations.

Ecophysiology traits

Morphological and ecophysiological characteristics are determinants of biomass productivity [4]. For that reason, independent leaf samples were obtained from the top of each tree and used for stable isotope analyses and leaf area estimation, respectively. This sort of measurements was carried considering their relationship with tree physiology and they are easy to collect (compared to physiological traits). Sampled leaves collected in August 2011 were processed to determine carbon (C) and nitrogen (N) concentrations and C and N stable isotope compositions by continuous flow isotope ratio mass spectrometry, at the UCDavis Stable Isotope Facility, as described elsewhere [5]. In this stage, the 461 clones were sampled. Additionally, a complimentary second subset of leaves, collected in August 2012, was utilized for measuring leaf area and dry biomass to estimate the specific leaf area (SLA) and N content per SLA ratio (NArea), according to Easlon *et al.* [27]. In this case, the subset of leaves was obtained from 177 clones, including three replicates per clone (one per block). These clones were chosen including those with extreme geographical origins and minimal and maximal volume indices. Thus, from these measurements, a set of variables were generated, including: C and N content, C:N ratio, carbon isotope discrimination (Δ), $\delta^{15}N$, SLA, and NArea.

DNA isolation

Additional young leaves were collected for DNA isolation. Circle punches were obtained and deposited in plastic vials, along with silica gel packets, until desiccation. This sort of leaf samples is compatible with automated grinding and pipetting systems. DNA isolation was performed using the Qiagen DNeasyPlant Mini Kit (Qiagen, Inc, Valencia, California, USA), according manufacturer's instructions. Quality was assessed with spectrophotometer (NanoDrop, Thermo Scientific). Acceptable extractions had a 260/280 ratio of 1.7-2.0 with a minimum concentration of 20 ng/µl.

Phytozome version 7.0 annotation and assembly files for *P. trichocarpa* (corresponding to assembly version 2.0 of the black cottonwood genome) were used to design oligonucleotide baits, complementary to short genomic regions that targeted exons, promoter, and intergenic control regions, as described by Zhou et al.[23]. A total of 230,720 baits of 120 bp were designed using SureSelect eArray software (Agilent Technologies, Santa Clara, California, USA). The baits targeted more than 39,000 of the 40,668 annotated protein-coding transcripts of the *P. trichocarpa* genome. As the cumulative length of the predicted exons exceeded the available baits, following bait design in eArray, we looped through the gene list, selecting one bait for each gene at each pass, until the maximum number of baits (i.e. 230,720) was reached. In addition to exons, baits were included in the design for genes with an annotated 5'-UTR targeting the 240 bp upstream, as well as 1000 baits targeting intergenic regions to be used as selectively neutral control regions. These control regions were selected at random from non-repetitive intergenic intervals at least 1000 bp from any gene model. This strategy of bait design has demonstrated previously that capture efficiency has not been significantly impacted by the presence of paralogous genes [25]. After design, a custom biotinylated RNA bait library was synthesized. Library preparation and target enrichment were performed following the Agilent SureSelect^{XT} protocol (Version B). Briefly, 3.0 µg of

poplar genomic DNA was sheared on a ultrasonicator (Covaris S220) at the Virginia Bioinformatics Institute (VBI), followed by end repair, 3'-end adenylation, adaptor ligation, and amplification. Agencourt AMPure XP beads were used to purify the libraries following each step, and library quality was assessed using an Agilent Bioanalyzer 2100 instrument, with Agilent DNA 1000 chips. Samples were randomly assigned one of 96 available index sequences and subsequently randomly assigned to groups of 16, each of which corresponding to a HiSeq sequencing lane. The prepared libraries were hybridized to the RNA baits in solution at 65 °C in an Eppendorf Mastercycler PCR machine (Eppendorf, Hamburg, Germany), and subsequently purified on magnetic beads. The multiplexed libraries were sequenced using an Illumina HiSeq 2500 System in a 2x100 paired-end format at VBI. Sequences of the prepared libraries are available at the Sequence Read Archive (http://www. ncbi.nlm.nih.gov/Traces/sra/) under the accession number SRA058855.

Data analysis and SNP calling

Short reads from poplar samples were pre-filtered using a collection of scripts in Biopieces (https://github.com/maasha/biopieces; version 2.0). First, interleaved pair-ended sequences were filtered based on the Illumina filter flag, and subsequently trimmed of adapters and bases with quality < 35. Following trimming, very short reads were eliminated (length < 35) to prevent ambiguous alignments. Lastly, reads having poor local quality scores (score < 25, window size = 5) were removed from the analysis. The short reads were aligned to the *Populus trichocarpa* (version 3.0) reference genome with the Burrows-Wheeler Aligner *mem* algorithm [28]. Resulting alignment files (in Sequence Alignment/Map, SAM, format) were converted via SAMtools v3.1 [29] to their binary versions (Binary Alignment/Map, BAM) for variant calling. Prior to SNP calling, duplicate reads were identified and removed using Picard software (version 2.6; https://broadinstitute.github.io/picard/command-line-overview.html) and the Genome Analysis Toolkit v3.x (GATK; https://software.broadinstitute.org/gatk/), with MarkDuplicates and DuplicateReadFilter functions, respectively [30]. Indels were realigned using the GATK IndelRealigner function. The HaplotypeCaller algorithm of GATK was then used to call SNPs (options: *min_base_quality_score > 9*, and *standard_min_confidence_threshold_for_calling > 29*) [31]. Variant calling was performed on individual chromosomes and scaffolds to reduce the run time. After merging all variant calling format (VCF) files, the VariantsToTable tool of GATK was used to produce SNP tables for downstream analysis. SNPs with a minor allele frequency < 5% and departure from Hardy-Weinberg equilibrium were excluded from the analyses. Similarly, SNPs that were missing data across more than 5% individuals (clones) were also removed from following stages. After these filtering steps, from 5.1 million identified SNPs, a final set comprised of 813,280 SNPs was utilized for GWAS.

GWAS analyses

The distributions of the different traits were checked for departures from normality. Logarithmic transformations were applied to normalize the variables Vol, DBF, C and N concentration, C/N ratio, SLA, NArea, lignin, C5-sugars and C6-sugars. Outlier observations were excluded from tests. Clonal means were adjusted by Best Linear Unbiased Predictor, using Proc MIXED in the software SAS v9.2 (SAS

Institute, Cary, NC, USA), in order to correct a significant block effect observed across the trial area. Prediction model included the factors clone and block, considered as random and fixed effects, respectively. The GWAS was based on Mixed Linear Models (MLM), implemented in the software GCTA v1.25 [32] (http://cnsgenomics.com/software/gcta/index.html). SNP data were first converted to PLINK format using TASSEL v3.0 [33] (http://www.maizegenetics.net/tassel), and then converted to binary PLINK (bed file) using PLINK v1.9 command line tool [34]. The genetic relatedness (kinship) was determined by the genetic relationship matrix (GRM) option at the GCTA's GRM module, based on identical-by-descent estimates [35]. The population structure matrix (Q) was estimated by a classical multidimensional scaling [36], also called Principal Coordinates Analysis, using the *cmdscale* function in the package *Stats* in R [37]. With the marker (M), kinship (K), population structure (Q) and predictor variables, we ran an "M+K+Q" linear mixed model to identify significant associations with traits. Marker and Q were assumed fixed effects, whereas K represented a random effect. The goodness of fit including (or not) Q was assessed for each trait by the Bayesian information criterion (BIC), utilizing the package *Stats* in R [37]. Associations were adjusted for multiple testing, using the Bonferroni correction, in the package *Stats* in R [37]. The threshold for p-values was 6.1479E-8 at significant level of 5% after Bonferroni correction (0.05/813,280). We additionally used the Random Forest method to estimate the percentage variance explained by the top SNPs for each trait (SNP contribution), by using the *randomForest* package in R [38]. The SNP contribution standard deviation (SCSD) was also calculated by the same package. We set the number of trees (*ntree*) grown as 2,000 and averaged 50 repeats per trait in our estimates.

Multiple-SNP testing for each trait was carried out using an overlapping sliding-window analysis on association results, based on the number of the significant associations (p-value < 0.00001) for 10 k window size with 1 k slide, using a custom script. Empirical p-values for each slide were estimated assuming the significant slides follow a Poisson distribution [39]. To this end, Poisson tests were carried out by comparing the rate of total significant SNPs and the total number of SNPs with the similar rate for the tested window using the probability formula below. (See Formula 2 in the Supplementary Files)

Where lambda (ג) is the average number of significant SNPs, *e* is the constant Euler's number and k is the average number of significant SNPs for the window being tested. Empirical p-values were recorded for each window per trait and then were adjusted by Bonferroni (0.05/number of tested windows) and reported for each trait. Finally, p-values were visualized using a Manhattan plot, highlighting only the significant slides.

Linkage disequilibrium

Linkage disequilibrium (LD) was estimated among pairwise combinations of SNPs per chromosome. It was expressed in terms of the squared correlation of allele frequencies r^2. The r^2 value between pairs of SNP markers, within each chromosome, was estimated using TASSEL 5.2 [33], utilizing the option of sliding window (120 SNPs per window). To assess the extent of LD, the decay of LD within physical

distance (base pairs) between SNPs, within each chromosome, was evaluated by nonlinear regression analysis of r^2 values [40]. Analysis was performed applying the NLIN procedure in SAS 9.4.

Gene models, annotations and expression data

Gene models and gene ontologies were obtained from the Phytozome platform version 12 (*Populus trichocarpa* v 3.0) and Quick GO site (http://www.ebi.ac.uk/QuickGO-Beta/), respectively.

Results And Discussion

We used GWAS to identify DNA polymorphisms associated with biomass production and wood chemical composition in *P. trichocarpa*, which determine its potential as feedstock for lignocellulosic ethanol. This approach complements our previous phenotypic characterization of the same association population [5] by identifying SNPs underlying traits of growth, ecophysiology and wood quality, the primary traits targeted for the development of genetically improved clones suitable for dedicated biomass and bioenergy plantations. An approach based on sequence capture allowed us to detect genotype-phenotype associations across the *P. trichocarpa* gene exome.

The association population used in this study consisted of 461 clones (from 101 provenances), comprising part of the natural distribution range of *P. trichocarpa* in the Pacific Northwest of the United States. In a previous study [5], we observed significant phenotypic and genetic variation for growth, spring bud phenology, water use efficiency, C and N assimilation, as well as lignocellulosic components and metabolome of wood (Table 1). Similarly, clonal repeatability, represented in terms of individual heritability estimates, also varied among the traits. We hypothesized from this information that multiple polymorphic loci across the genome should be detected in association with phenotypes, and particularly, those with high heritability should reveal a large number of significant SNP-markers.

Genotyping

The processes of exome sequencing and genotyping identified 5.1 million SNPs across the *P. trichocarpa* genome in the association population, and after filtering, a set of 813,280 SNPs was used for association analyses (Table 2). The number of selected SNPs was proportional to chromosome size, ranging from 29,287 to 100,299 SNPs, for chromosomes 9 and 1, respectively (Table 2, Fig. 1a). Considering the full genome length, an average of one SNP every 482 bp (Table 2) was included in the analyses. Taking advantage of the full genome assembly, genotyping methodologies such as those based on sequence capture can target entire exons or genes across the genome, avoiding bias arising by a priori selection of candidate loci [23, 25]. In comparison to similar preceding studies that used SNP array platforms [6, 18, 19, 41], the number of SNPs in our analyses represent an increase in the power of applied genomic scanning. However, this amount is lower than the utilized by approaches based on whole-genome sequencing developed recently [7, 15].

Intra-chromosomal linkage disequilibrium

The extent of linkage disequilibrium (LD) was analyzed across each chromosome. On average, the LD over physical distance decayed below r^2 0.2 at 26.9 kbp. A representative example, for Chromosome 12, is depicted in Fig. 1 b. The complete set of chromosomes with its LD is included in Fig. S1. The decay varied depending on specific chromosomes, with the most rapid decay observed on chromosomes 7 and 15 (r^2 0.2 at 18.9 kbp) and the slowest decay on chromosome 11 (r^2 0.2 at 51.6 kbp). Genome-wide LD decay exhibited different extents among chromosomes (Table 2). LD decay to r^2<0.2 was observed on average at 26.9 kbp. High variation of LD across the genome (among and within chromosomes) has been reported for this species [23]. The estimated extent of LD decay predicted in our study is higher than the observed by Wegrzyn et al. [18] (r^2 0.2 at ~0.5 kbp) and Wang et al.[42] (r^2 0.2 at ~8 kbp) for *P. trichocarpa.* Distinct methodologies, number of markers, population sizes, genetic origins and standard errors among the studies may account for the different findings. Compared with other tree species extent of LD estimated in this study is similar to species belonging to *Fraxinus* [43], *Prunus* [44] and *Eucalyptus* [45] genus.

Single SNP-marker associations

Significant associations (p-value < 6.1479E-8) were identified for DBH, h, leaf C and N content, and $\delta^{15}N$. Figure 2 (a and c) depicts the number of associations detected per chromosome for a selected set of traits. A detailed list for each trait is provided in Table S1. Similarly, Manhattan plots for each phenotype are included in Fig. S2.

In general, and consistently with chromosome length, the highest numbers of significant associations were observed for chromosomes 1 and 5. The lowest number of associations was observed for chromosome 16. The proportion of significant SNPs of the total analyzed, ranged from 0.02 ‰ to 0.50 ‰ for leaf C content on chromosome 10, along with $\delta^{15}N$ on chromosomes 6 and 10, and leaf N content on chromosome 5 (Table S1b), respectively. In the case of growth traits, 2 and 148 associations were detected for DBH and h, respectively. Within the ecophysiological traits, the number of significant associations ranged from 12 to 220 for C content and leaf N-content, respectively. For traits related to the chemical composition of wood, associated SNP-markers were over the significance cutoff (p-value < 6.1479E-8). Similarly, in the case of wood metabolites, considering a selected subset of those with the top five highest heritability estimates, no significant associations meeting the adjusted p-value were identified for Adenosine (Ade), Hydroxybenzoic Acid (HbA), Galactinol (Gal), Galactonic Acid (GAc) and Alpha tocopherol (Toc). The proportion of phenotypic variation accounted for the cumulative effect of significantly associated SNPs was 0.2 %, 1.1 %, 0.1 %, 0.7 % and 0.7 % for DBH, h, leaf C content, leaf N content, and $\delta^{15}N$, respectively.

Significant single nucleotide polymorphisms associated with phenotype were identified mostly in exonic regions. SNPs are part of genes encoding proteins belonging to the functional classes: Protein Synthesis/Modification (54.5 %), DNA/RNA Metabolism (27.3 %), Energy/Metabolism (9.1 %) and Signal transduction (9.1 %) (Fig. 3a). A list with these SNPs and genes is given in Table S3. An example for the Protein Synthesis/Modification category was a gene encoding a Periodic Tryptophan Protein 1

(Potri.007G019500), which was associated with height, and leaf N and $\delta^{15}N$. Among genes related with proteins involved in DNA/RNA Metabolism, one for a helicase senataxin (without gene model in Phytozome) was significant for height and leaf N. For genes in the Energy/Metabolism functional class, a representative was one (Potri.015G119700) encoding a Domain of unknown function (PGG), which was associated with DBH. For the Signal transduction class, the gene encoding a Rop Guanine Nucleotide Exchange Factor 1 (Potri.009G140100) was significant for height and leaf N.

Considering the applied significance threshold with Bonferroni correction (p-value < 6.1479E-8), GWAS performed on single-SNPs was successful in identifying polymorphisms associated with growth traits (DBH and h), leaf C and N-contents, as well as stable isotope parameters ($\delta^{15}N$) (Fig 2, Table S1). For traits related to spring bud phenology (DBF), wood chemical components (C5 and C6 sugars, lignin) and wood metabolites (GAc, Gal and HbA) significant associations at p-value< 0.0001 were detected, but they did not reach the adjusted threshold. The presence or lack of significant SNPs for these traits appears to be independent of heritability estimates for each. For some traits with moderate to high H^2_i (e.g. S:G ratio or DBF), GWAS did not detect single-SNP associations. On the other hand, for traits with low to moderate H^2_i (e.g. leaf C-content and $\delta^{15}N$) a relatively higher number of SNPs were identified. Similar situations were observed for phenology traits in previous studies with *P. trichocarpa* [19]. On average for all traits with significant associations ~ 1 % of phenotypic variation was accounted for by the cumulative effect of significant SNPs. The influence of multiple SNPs associated with phenotypes is particularly interesting in the context of the development of models for genomic selection, where large numbers of markers are utilized to predict the genetic merit of individuals [46]. Differences among traits in terms of the number of significant SNP-markers suggest the differential effect of both the variable number of SNPs influencing each trait and the individual impact of some SNPs. In that sense, some individual SNPs could have a such low effect size that none reach statistical significance. Furthermore, the apparent lack of correspondence between estimates of H^2_i and the phenotypic variance collectively accounted for by SNPs, could be explained by non-additive effects (e.g epistasis, GxE effect) or epigenetic factors acting on some traits. These types of effects are usually underestimated because MLM utilized for GWAS only suppose additive interactions [19]. Finally, another factor influencing the number of significant associated SNPs (and their effect on phenotypes) deals with the complexity of analyzing thousands of single markers across the genome. Stringent thresholds for controlling type I error are required for p-value adjustment in GWAS, given the correlated nature of markers along a chromosome [47]. For example, it has been suggested that the general applicability of the traditional false discovery ratio (FDR) [48] may suffer from several problems when applied to association analysis of a single trait [49]. In that sense, we utilized the Bonferroni correction to define the significance threshold. Thus, in spite of significant associations were detected at p-value < 0.00001 (and even lesser) in traits such as Vol, DBF, lignin or GAc, they did not reach the adjusted p-value threshold and were considered non-significant.

Sliding window analyses

The multiple-marker analysis by sliding-window allowed us to identify genomic regions containing different sets of SNPs jointly associated with each trait. Figure 4a depicts a representative Manhattan plot with the significant windows identified for leaf $\delta^{15}N$. Manhattan plots for other traits are included in Fig. S3. A variable number of windows per chromosome were detected among the phenotypes (Fig. 2 b and d). The total number of significant windows ranged from 6 for HbA, to 192 for N content (Table S2). For most traits, the main contributions were observed by chromosomes 8 and 1. However, for traits such as DBF, C:N, $\delta^{15}N$, and Toc, the most relevant chromosomes in terms of the number of significant windows included to 6, 4, 5 and 10, respectively. The multiple-SNP approach applied by sliding window analysis has been proposed as a robust alternative for identifying clustered significant patterns of SNPs, that are associated with complex traits, in a chromosomal context in humans and plants [39, 50-52]. In our study, significant windows identified a series of SNP clusters which were coincident with coding regions of multiple genes (Table S4). The graphical relationship between SNPs identified by single-marker associations and the detection by sliding window analysis is depicted in Fig.4, where the highlighted window (Fig.4 a) contains 14 significant SNPs belonging to the *XRN4* gene (Fig.4 b). Additionally, information coming from both detection approaches allowed us to define genome zones with high LD, significantly associated with phenotypic variation, revealing the presence of phenotypically-relevant haplotypes (Fig. 4c). Although more evidence will be necessary, haplotype blocks defined by this way could be indicative of polymorphic regions with pleiotropic effects.

Considering the top three most significant windows across all chromosomes and traits, the most represented functional classes were Protein with Unknown Function (21.2 %), Energy/Metabolism (21.1 %), Protein Synthesis/Modification (19.2 %), and Transcription (15.4 %) (Fig. 3b). A list with the windows and genes included in these classes are given in Table S4. Some of the detected genes encoding proteins with roles in Protein Synthesis/Modification were those expressing Similar to Threonyl-tRNA Synthetase (Potri.008G145600), Interleukin-1 Receptor-Associated Kinase 4 (Potri.008G145900) and Leucine Rich Repeat (LRR1) (Potri.005G015700) associated with wood C5-sugars, C6-sugars and height, respectively. An example of the genes dealing with proteins belonging to the Energy/Metabolism class were Exostosin Heparan Sulfate Glycosyltransferase-Related (Potri.010G197900) and Similar to Aldehyde Dehydrogenase 1 Precursor (Potri.012G078700), associated with $\delta^{15}N$ and DBF, respectively. Among genes encoding enzymes involved in Transcription, Similar to Agamous-like MADS Box Protein AGL12 (Potri.019G076800) and WRKY Transcription Factor 10-related (Potri.013G086000), were associated with Gal and C5-sugars, respectively. Concerning the Protein Synthesis/Modification class, examples of identified genes are those encoding a Similar to Threonyl-tRNA synthetase and Leucine Rich Repeat (LRR_1)//Leucine rich repeat (LRR_8), associated with C5-sugars and height, respectively.

Genes detected by single-SNP association and sliding windows approaches

We also verified the consistency between SNP-markers identified by single-SNP association and sliding window analyses. To this end, we considered as an example the most significant window (#154; p-value 4.70E-25) detected in chromosome 5 for leaf $\delta^{15}N$ (Fig. 4a). The SNPs identified within the window were part of a gene (Potri.005G048900) encoding a Similar to 5'-3' Exoribonuclease (XRN4) Gene, which is

involved in disease resistance, response to ethylene, RNAi, and miRNA-mediated RNA decay [53]. The associated window comprised 64 SNPs (Fig. 4b). Fourteen of these markers, were also identified by the single-marker association, indicating the consistency between both approaches. Particularly, markers such as S05_3547832, S05_3547864, S05_3547904 and S05_3548573 were in high LD ($r^2 > 0.75$) and produced significant variation at the level of leaf $\delta^{15}N$ means (Fig. 4c). Alternatively, alleles for those SNPs represent intronic and non-synonymous polymorphisms, involving possible effects on transcript splicing, and protein structure and function.

Significant associations for traits underlying growth, nutrient metabolism and xylem formation, among others, define SNPs and genes, which might represent logical candidates for functional studies focused on confirming their role and impact on phenotype of *Populus* species. Considering their high significance and the simultaneous detection by the single-SNP association and/or sliding windows approaches, we centered part of our analysis on gene Similar to 5'-3' Exoribonuclease (*XRN4*), which included SNPs and windows associated with leaf $\delta^{15}N$. The exoribonuclease function of this gene links it to transcription, RNA metabolism and RNA interference in eukaryotes [54]. In plants, it has been related to ethylene signaling [55] and response of plants to abiotic and biotic stresses [56, 57]. Mutation of members of the *XNR* gene family produced sensitivity to N starvation in *Saccharomyces cerevisiae* [58] and morphological alterations in *Arabidopsis thaliana* [59, 60]. Thus, association of *XRN4* with leaf $\delta^{15}N$, an indicator of N use efficiency [61], could be related to the N metabolism and mobilization at leaves, particularly during the last third of the growing season, when sampling was done. More studies will be required to detect possible effects of SNPs at *XRN4* on photosynthesis and biomass production.

Chemical composition of wood

Association analyses of the chemical composition of *P. trichocarpa* wood was previously performed in the same population by our group [18] using a candidate gene approach. We carried out a comparison between the results from both studies considering the 40 candidate genes utilized by Wegrzyn *et al.* [18], which encodes enzymes from the cellulose and lignin biosynthesis pathways and cytoskeletal proteins. Association results in the present study were significant only under a p-value < 0.0001 threshold. Results indicated both overlap and divergence between the two studies (Fig. 5 and Table S5). SNPs within genes encoding cellulose synthase (*CesA1A*) were significantly associated with C6-sugars in both studies. For this trait, we also detected SNPs in *TUA5* gene, which were not identified by Wegrzyn *et al.* [18]. In the case of lignin content, members of the cellulose synthase gene family (*CesA2B* and *CesA1A*) were differentially detected (differentially) in both studies. We also identified SNPs belonging to 4- Serine Hydroxymethyl Transferase SHMT6. Finally, for the S:G ratio, our analyses detected SNPs in Laccase *LAC1A*, Phenylalanine Ammonia-Lyase *PAL5* genes, which were not identified by Wegrzyn *et al.* (2010). In spite of the genotypes and wood chemical characterization methods were mostly the same in both studies, distinct trial sites (the first in Westport, Oregon, and the second in Davis, California), sampling height or differential presence of juvenile wood, among other factors, might explain the differences in the findings reported previously [18] and those described in the present work.

Conclusions

Forest trees are an important source for multiple wood and non-wood products. Genetic improvement aimed to develop such products, including lignocellulosic biofuels, depends on the variation underlying commercial traits. This variation is characterized by a complex genetic control and the influence of environmental factors. In this study, we identified a series of DNA polymorphisms controlling the phenotypic variation of growth, nitrogen use and wood composition of *P. trichocarpa* clones at different levels. Our results thus provide a starting point to define candidate genes suitable for functional characterizations addressed to confirm their biological role. At the same time, the genome-wide scale applied for the association analyses revealed a large number of SNPs which could be utilized to develop genomic selection schemes. Further efforts to define the utility of SNP polymorphisms in generating genomic breeding values will illuminate the path to breeding programs that incorporate molecular markers for bioethanol production. The upshot will be the estimation of parental hybridization values and an earlier, more precise genotypic selection from segregating F1 hybrid populations, that together will increase the overall magnitude of the realized genetic gain per unit of time.

Abbreviations

Ade: Adenosine; C:N: Leaf C:N ratio; C5: Wood 5-carbon sugars; C6: Wood 6-carbon sugars; DBF: Days to bud flush; DBH: Diameter; GAc: Galactonic acid; Gal: Galactinol; GWAS: Genome wide association study; h: Height; H^2_i: Individual broad-sense heritability; HbA: 4-Hydroxybenzoic acid; LD: Linkage disequilibrium; Narea: N content:SLA ratio; S:G: Wood syringil:guayacil ratio; SLA: Specific leaf area; SNP: Single Nucleotide Polymorphism; Toc: Alpha tocopherol; Vol: Volume index; Δ: carbon isotope discrimination.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Availability of data and materials

Sequences of the prepared libraries for exome capture are available at the Sequence Read Archive (http://www. ncbi.nlm.nih.gov/Traces/sra/) under the accession number SRA058855. Data used in the phenotypic characterization is available at the TreeGenes platform (http://dendrome.ucdavis.edu/treegenes/) under the accession number TGDR050.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was funded by the Advanced Hardwood Biofuels Northwest Project, supported by Agriculture and Food Research Initiative Competitive (Grant no. 2011-68005-30407, USDA National Institute of Food and Agriculture) and by the National Science Foundation Plant Genome Research Program (IOS Grant no. 1054444 to JAH).

Author's contributions

B.J.S., D.B.N and F.P.G. planned and designed the research. F.P.G and H.S. analysed data. D.B.N, F.P.G, J.H. and H.S. interpreted data. F.P.G, J.H.R and R.S. conducted fieldwork and data and sample collection. J.H.R, M.D., O.F., R.F. and R.S. processed and analyzed samples. B.J.S., D.B.N, F.P.G, J.H., J.H.R, H.S. and R.F. wrote the manuscript.

Acknowledgements

We would like to thank the California Agricultural Experiment Station for the provided support and two anonymous reviewers for their helpful comments on earlier version of the manuscript.

Author details

[1]Department of Plant Sciences, University of California at Davis, CA 95616, USA. [2] Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. [3] Department of Land, Air and Water Resources, University of California, Davis, CA 95616, USA. [4] Department of Molecular and Cellular Biology & Genome Center, University of California, Davis, CA 95616, USA. [5] Biological Research Group, GreenWood Resources, Portland, OR 97201, USA. [6] National Renewable Energy Laboratory, Golden, CO 80401, USA. [7]Bioenergy Research Center, University of California at Davis, Davis, CA 95616, USA. [8]Instituto de Ciencias Biológicas, Universidad de Talca, P.O. Box 747, Chile.

References

1. Porth I, El-Kassaby YA: **Using Populus as a lignocellulosic feedstock for bioethanol**. *Biotechnology Journal* 2015, **10**(4):510-524.
2. Davis JM: **Genetic Improvement of Poplar (*Populus* spp.) as a Bioenergy Crop**. In: *Genetic Improvement of Bioenergy Crops.* Edited by Vermerris W: Springer New York; 2008: 397-419.
3. Stanton BJ, Neale D, Li S: ***Populus* breeding: from the classical to the genomic approach**. In: *Genetics and genomics of Populus.* Edited by Jansson S, Bhalerao R, Groover A, vol. 8. New York: Springer; 2010: 309-348.

4. Mitchell CP: **Ecophysiology of short rotation forest crops**. *Biomass and Bioenergy* 1992, **2**(1–6):25-37.
5. Guerra F, Richards J, Fiehn O, Famula R, Stanton B, Shuren R, Sykes R, Davis M, Neale D: **Analysis of the genetic variation in growth, ecophysiology, and chemical and metabolomic composition of wood of Populus trichocarpa provenances**. *Tree Genetics & Genomes* 2016, **12**(1):1-16.
6. McKown AD, Guy RD, Klápště J, Geraldes A, Friedmann M, Cronk QCB, El-Kassaby YA, Mansfield SD, Douglas CJ: **Geographical and environmental gradients shape phenotypic trait variation and genetic structure in *Populus trichocarpa***. *New Phytologist* 2014, **201**(4):1263-1276.
7. Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G *et al*: **Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations**. *Nat Genet* 2014, **46**(10):1089-1096.
8. Porth I, Klápště J, Skyba O, Friedmann MC, Hannemann J, Ehlting J, El-Kassaby YA, Mansfield SD, Douglas CJ: **Network analysis reveals the relationship among wood properties, gene expression levels and genotypes of natural Populus trichocarpa accessions**. *New Phytologist* 2013, **200**(3):727-742.
9. Porth I, Klapšte J, Skyba O, Hannemann J, McKown AD, Guy RD, DiFazio SP, Muchero W, Ranjan P, Tuskan GA *et al*: **Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms**. *New Phytologist* 2013, **200**(3):710-726.
10. Porth I, Klápště J, Skyba O, Lai BSK, Geraldes A, Muchero W, Tuskan GA, Douglas CJ, El-Kassaby YA, Mansfield SD: **Populus trichocarpa cell wall chemistry and ultrastructure trait variation, genetic control and genetic correlations**. *New Phytologist* 2013, **197**(3):777-790.
11. Scaracia-Mugnozza GE, Ceulemans R, Heilman PE, Isebrands JG, Stettler RF, Hinckley TM: **Production physiology and morphology of *Populus* species and their hybrids grown under short rotation. II. Biomass components and harvest index of hybrid and parental species clones**. *Canadian Journal of Forest Research* 1997, **27**(3):285-294.
12. Zabek LM, Prescott CE: **Biomass equations and carbon content of aboveground leafless biomass of hybrid poplar in Coastal British Columbia**. *Forest Ecology and Management* 2006, **223**(1–3):291-302.
13. Bradshaw HD, Stettler RF: **Molecular genetics of growth and development in *Populus*. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree**. *Genetics* 1995, **139**(2):963-973.
14. McKown A, Klapste J, Guy RD, El-Kassaby YA, Mansfield SD: **Ecological genomics of variation in bud-break phenology and mechanisms of response to climate warming in Populus trichocarpa**. *New Phytologist* 2018, **220**(1):300-316.
15. Chhetri HB, Macaya-Sanz D, Kainer D, Biswal AK, Evans LM, Chen J-G, Collins C, Hunt K, Mohanty SS, Rosenstiel T *et al*: **Multitrait genome-wide association analysis of Populus trichocarpa identifies key polymorphisms controlling morphological and physiological traits**. *New Phytologist* 2019, **223**(1):293-309.

16. McKown AD, Guy RD, Quamme L, Klápště J, La Mantia J, Constabel CP, El-Kassaby YA, Hamelin RC, Zifkin M, Azam MS: **Association genetics, geography and ecophysiology link stomatal patterning in Populus trichocarpa with carbon gain and disease resistance trade-offs**. *Mol Ecol* 2014, **23**(23):5771-5790.
17. Monclus R, Villar M, Barbaroux C, Bastien C, Fichot R, Delmotte FM, Delay D, Petit JM, Brechet C, Dreyer E *et al*: **Productivity, water-use efficiency and tolerance to moderate water deficit correlate in 33 poplar genotypes from a *Populus deltoides x Populus trichocarpa* F1 progeny**. *Tree Physiology* 2009, **29**(11):1329-1339.
18. Wegrzyn JL, Eckert AJ, Choi M, Lee JM, Stanton BJ, Sykes R, Davis MF, Tsai C-J, Neale DB: **Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem**. *New Phytologist* 2010, **188**(2):515-532.
19. McKown AD, Klápště J, Guy RD, Geraldes A, Porth I, Hannemann J, Friedmann M, Muchero W, Tuskan GA, Ehlting J *et al*: **Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of Populus trichocarpa**. *New Phytologist* 2014, **203**(2):535-553.
20. Guerra FP, Wegrzyn JL, Sykes R, Davis MF, Stanton BJ, Neale DB: **Association genetics of chemical wood properties in black poplar (*Populus nigra*)**. *New Phytologist* 2013, **197**(1):162-176.
21. Fahrenkrog AM, Neves LG, Resende MF, Jr., Vazquez AI, de Los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB *et al*: **Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides**. *The New phytologist* 2017, **213**(2):799-811.
22. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C *et al*: **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing**. *Nat Biotech* 2009, **27**(2):182-189.
23. Zhou L, Bawa R, Holliday JA: **Exome resequencing reveals signatures of demographic and adaptive processes across the genome and range of black cottonwood (Populus trichocarpa)**. *Mol Ecol* 2014, **23**(10):2486-2499.
24. Kaur P, Gaikwad K: **From Genomes to GENE-omes: Exome Sequencing Concept and Applications in Crop Improvement**. *Frontiers in plant science* 2017, **8**:2164-2164.
25. Zhou L, Holliday JA: **Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture**. *BMC Genomics* 2012, **13**:703-703.
26. Holliday JA, Zhou L, Bawa R, Zhang M, Oubida RW: **Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in Populus trichocarpa**. *New Phytologist* 2016, **209**(3):1240-1251.
27. Easlon HM, Nemali KS, Richards JH, Hanson DT, Juenger TE, McKay JK: **The physiological basis for genetic variation in water use efficiency and carbon isotope composition in *Arabidopsis thaliana***. *Photosynthesis Research* 2014, **119**(1-2):119-129.
28. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
30. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M *et al*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nat Genet* 2011, **43**(5):491-498.
31. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.
32. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: a tool for genome-wide complex trait analysis**. *Am J Hum Genet* 2011, **88**(1):76-82.
33. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES: **TASSEL: software for association mapping of complex traits in diverse samples**. *Bioinformatics* 2007, **23**(19):2633-2635.
34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**(3):559-575.
35. Zheng X, Weir BS: **Eigenanalysis of SNP data with an identity by descent interpretation**. *Theor Popul Biol* 2016, **107**:65-76.
36. Gower JC: **Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis**. *Biometrika* 1966, **53**(3/4):325-338.
37. R-Core-Team: **R: A language and environment for statistical computing**. In. Vienna, Austria: R Foundation for Statistical Computing; 2014.
38. Liaw A, Wiener M: **Classification and Regression by Random Forest**. *R News* 2002, **2**:18-22.
39. Sun YV, Jacobsen DM, Turner ST, Boerwinkle E, Kardia SLR: **Fast implementation of a scan statistic for identifying chromosomal patterns of genome wide association studies**. *Computational Statistics & Data Analysis* 2009, **53**(5):1794-1801.
40. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Structure of linkage disequilibrium and phenotypic associations in the maize genome**. *Proceedings of the National Academy of Sciences* 2001, **98**(20):11479-11484.
41. Muchero W, Guo J, DiFazio SP, Chen J-G, Ranjan P, Slavov GT, Gunter LE, Jawdy S, Bryan AC, Sykes R *et al*: **High-resolution genetic mapping of allelic variants associated with cell wall chemistry in Populus**. *BMC Genomics* 2015, **16**(1):24.
42. Wang J, Street NR, Scofield DG, Ingvarsson PK: **Natural Selection and Recombination Rate Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related Populus Species**. *Genetics* 2016, **202**(3):1185-1200.
43. Sollars ESA, Harper AL, Kelly LJ, Sambles CM, Ramirez-Gonzalez RH, Swarbreck D, Kaithakottil G, Cooper ED, Uauy C, Havlickova L *et al*: **Genome sequence and genetic diversity of European ash trees**. *Nature* 2016, **541**:212.

44. Campoy JA, Lerigoleur-Balsemin E, Christmann H, Beauvieux R, Girollet N, Quero-García J, Dirlewanger E, Barreneche T: **Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of Prunus avium L. landraces and bred cultivars**. *BMC Plant Biology* 2016, **16**(1):49.
45. Müller BSF, Neves LG, de Almeida Filho JE, Resende MFR, Muñoz PR, dos Santos PET, Filho EP, Kirst M, Grattapaglia D: **Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of Eucalyptus**. *BMC Genomics* 2017, **18**:524.
46. Isik F: **Genomic selection in forest tree breeding: the concept and an outlook to the future**. *New Forests* 2014, **45**(3):379-401.
47. Balint-Kurti P, Simmons SJ, Blum JE, Ballaré CL, Stapleton AE: **Maize Leaf Epiphytic Bacteria Diversity Patterns Are Genetically Correlated with Resistance to Fungal Pathogen Infection**. *Molecular Plant-Microbe Interactions* 2010, **23**(4):473-484.
48. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency**. *The Annals of Statistics* 2001:1165-1188.
49. Chen L, Storey JD: **Relaxed Significance Criteria for Linkage Analysis**. *Genetics* 2006, **173**(4):2371-2381.
50. Sun YV, Levin AM, Boerwinkle E, Robertson H, Kardia SLR: **A scan statistic for identifying chromosomal patterns of SNP association**. *Genetic Epidemiology* 2006, **30**(7):627-635.
51. Asimit JL, Andrulis IL, Bull SB: **Regression models, scan statistics and reappearance probabilities to detect regions of association between gene expression and copy number**. *Statistics in Medicine* 2011, **30**(10):1157-1178.
52. Morrison KM, Simmons SJ, Stapleton AE: **Loci controlling nitrate reductase activity in maize: ultraviolet-B signaling in aerial tissues increases nitrate reductase activity in leaf and root when responsive alleles are present**. *Physiologia Plantarum* 2010, **140**(4):334-341.
53. Rymarquis LA, Souret FF, Green PJ: **Evidence that XRN4, an Arabidopsis homolog of exoribonuclease XRN1, preferentially impacts transcripts with certain sequences or in particular functional categories**. *RNA* 2011, **17**(3):501-511.
54. Chang JH, Xiang S, Xiang K, Manley JL, Tong L: **Structural and biochemical studies of the 5'→3' exoribonuclease Xrn1**. *Nature Structural &Amp; Molecular Biology* 2011, **18**:270.
55. Potuschak T, Vansiri A, Binder BM, Lechner E, Vierstra RD, Genschik P: **The Exoribonuclease XRN4 Is a Component of the Ethylene Response Pathway in Arabidopsis**. *The Plant Cell* 2006, **18**(11):3047-3057.
56. Merret R, Descombin J, Juan Y-t, Favory J-J, Carpentier M-C, Chaparro C, Charng Y-y, Deragon J-M, Bousquet-Antonelli C: **XRN4 and LARP1 Are Required for a Heat-Triggered mRNA Decay Pathway Involved in Plant Acclimation and Survival during Thermal Stress**. *Cell Reports* 2013, **5**(5):1279-1293.
57. Rymarquis L, Souret F, Green P: **Evidence that XRN4, an Arabidopsis homolog of exoribonuclease XRN1, preferentially impacts transcripts with certain sequences or in particular functional categories**.

Rna 2011, **17**(3):501-511.

58. Sinturel F, Bréchemier-Baey D, Kiledjian M, Condon C, Bénard L: **Activation of 5'-3' exoribonuclease Xrn1 by cofactor Dcs1 is essential for mitochondrial function in yeast**. *Proceedings of the National Academy of Sciences* 2012, **109**(21):8264-8269.
59. Kim B-H, Von Arnim AG: **FIERY1 regulates light-mediated repression of cell elongation and flowering time via its 3'(2'),5'-bisphosphate nucleotidase activity**. *The Plant Journal* 2009, **58**(2):208-219.
60. Hirsch J, Misson J, Crisp PA, David P, Bayle V, Estavillo GM, Javot H, Chiarenza S, Mallory AC, Maizel A *et al*: **A Novel fry1 Allele Reveals the Existence of a Mutant Phenotype Unrelated to 5'->3' Exoribonuclease (XRN) Activities in Arabidopsis thaliana Roots**. *PLOS ONE* 2011, **6**(2):e16724.
61. Cernusak LA, Winter K, Turner BL: **Plant delta 15N correlates with the transpiration efficiency of nitrogen acquisition in tropical trees**. *Plant physiology* 2009, **151**(3):1667-1676.

Tables

Table 1. Summary statistics for traits studied in the *Populus trichocarpa* association population. Columns “Mean”, “Std. Dev.”, “C.V” and “” were extracted from Guerra et al. [5].

Trait		Unit	Mean	Std. Dev.	C.V. (%)	H^2_c
Growth	Diameter (DBH)	mm	53.2	7.9	14.8	0.52
	Height (h)	dm	67.1	4.1	6.1	0.42
	Volume index (Vol)	m^3	0.016	0.005	31.3	0.53
Phenology	Days to bud flush (DBF)	Julian days	87.4	7.8	8.9	0.9
Ecophysiology	Leaf C content	% DW	44.4	1.6	3.6	0.09
	Leaf N content	% DW	3.2	0.3	9.4	0.28
	Leaf C:N ratio (C:N)	kg C/kg N	14.2	1.3	9.2	0.33
	Leaf Δ	‰	19.2	0.7	3.6	0.26
	Leaf δ ^{15}N	‰	2.5	0.4	16.0	0.25
	Specific leaf area (SLA)	m^2/kg DW	12	1.5	12.5	0.27
	N content : SLA ratio (NArea)	g N/m^2	2.8	0.4	14.3	0.28
Wood chem. components	Wood 5-carbon sugars (C5)	%	36	2.2	6.1	0.07
	Wood 6-carbon sugars (C6)	%	42.3	3.3	7.8	0.08
	Wood lignin	%	22.7	1	4.4	0.15
	Wood syringil:guayacil ratio (S:G)	fold	1.9	0.1	5.3	0.58
Wood metabolites	Galactonic acid (GAc)	R.A.[a]	0.6	0.4	62.8	0.22
	Galactinol (Gal)	R.A.[a]	144.1	75.0	52.0	0.28
	Alpha tocopherol (Toc)	R.A.[a]	69.3	31.1	44.9	0.16
	Adenosine (Ade)	R.A.[a]	2.8	1.0	33.4	0.25
	4-Hydroxybenzoic acid (HbA)	R.A.[a]	5.9	4.6	78.3	0.45

Table 2. Summary of amount of analyzed SNP markers and intrachromosomal LD decay across the *Populus trichocarpa* genome. Linkage disequilibrium decay is referred to the

physical distance (kbp) where LD = 0.2

Chr.	Size (Mbp)	Analyzed SNPs	Frequency (bp/SNP)	LD Decay[a] (kbp)
1	50.5	100,299	503.4	29.99
2	25.3	47,563	531.1	27.49
3	21.8	49,962	436.7	27.19
4	24.3	47,671	509.1	22.36
5	25.9	52,236	495.6	23.35
6	27.9	49,374	565.3	27.21
7	15.6	30,295	515.3	18.85
8	19.5	43,099	451.6	21.99
9	12.9	29,287	442.1	21.63
10	22.6	46,758	482.9	24.21
11	18.5	38,563	479.8	51.63
12	15.8	31,964	493.1	25.13
13	16.3	30,493	535.2	28.07
14	18.9	40,482	467.4	29.65
15	15.3	33,418	457.2	18.85
16	14.5	32,006	452.9	26.22
17	16.1	39,114	411.1	33.83
18	17	34,049	498.1	33.39
19	15.9	36,647	435.0	19.26
Total	394.5	813,280	-	-
		Mean	482.3	26.86

Additional File Legends

Additional File 1: Table S1a. Number of significant single SNP-markers.

Additional File 1: Table S1b. Proportion of significant SNPs on total analyzed SNPs.

Additional File 1: Table S2. Number of significant sliding windows .

Additional File 1: Table S3. Top three significantly associated single SNPs.

Additional File 1: Table S4. Top three significantly associated SNP-windows.

Additional File 2: Table S5. SNPs belonging to genes encoding enzymes for lignin and cellulose biosynthesis pathways.

Additional File 3: Figure S1. Linkage disequilibrium decay per chromosome.

Additional File 4: Figure S2. Manhattan plots for assessed traits.

Additional File 5: Figure S3. Manhattan plots for sliding window analysis tests.

Figures

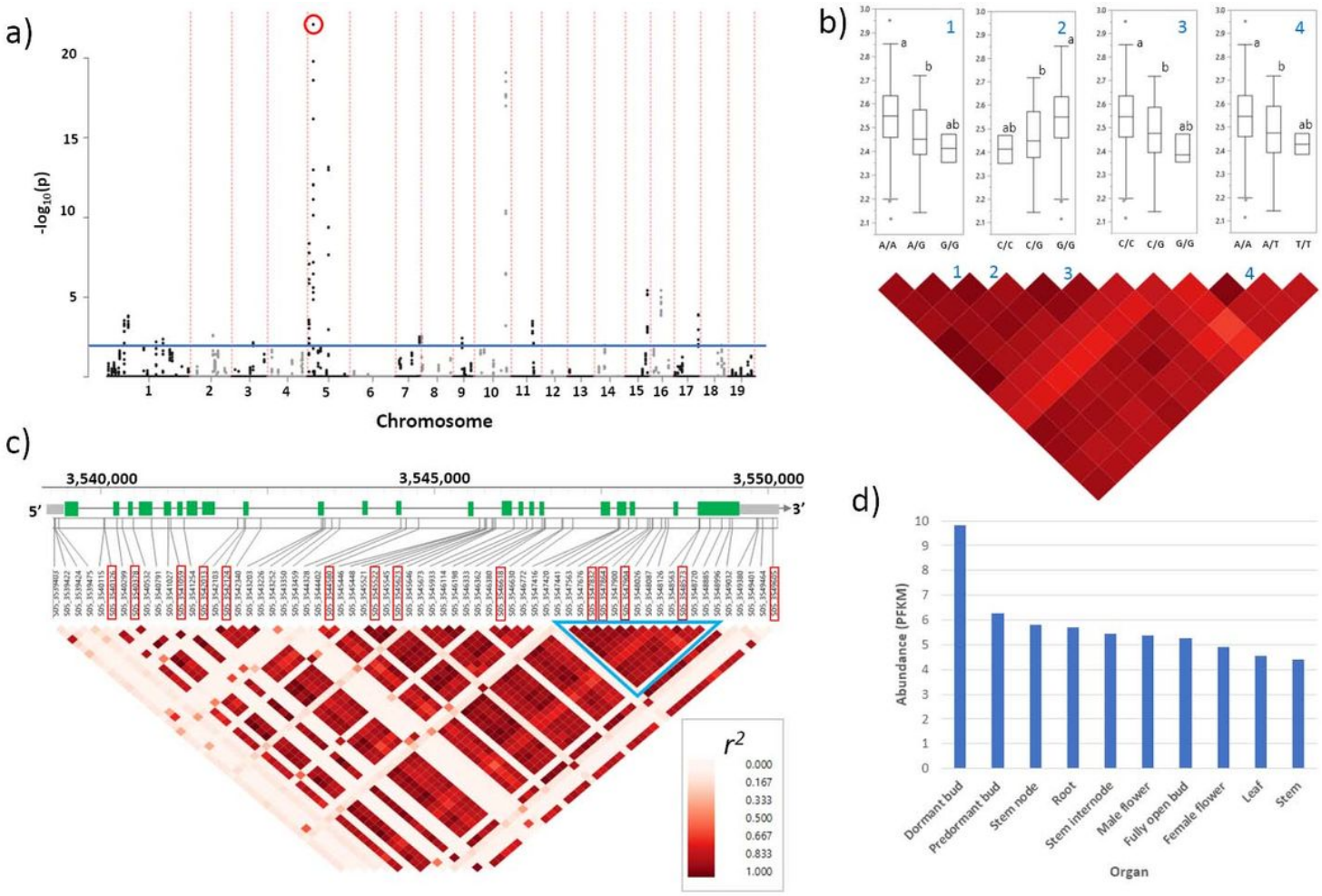

Figure 0

Detailed characterization of Similar to 5'-3' Exoribonuclease (XRN4) gene (Potri.005G048900) associated with leaf δ15N. a) Manhattan plot for leaf δ15N highlighting (red circle) the window containing significant SNPs for the gene. The horizontal blue line indicates a -log10 (p-value) of 2 (equivalent to p-value = 0.01). b) LD heat map for the analyzed SNPs located at gene. Red bars at the top correspond to SNPs identified as significantly associated with δ15N by single-marker association tests. c) Detailed view for the blue light triangle depicted in b). Numbers 1, 2, 3 and 4 are the markers S05_3547832, S05_3547864, S05_3547904 and S05_3548573, respectively. Boxplots shows the effects of genotypes on leaf δ15N. Different letters indicate significant differences among adjusted means (Tukey's HSD test; α=0.001). d) Referential gene expression profile obtained from available data for Potri.005G048900 at Phytozome-Phytomine platform.

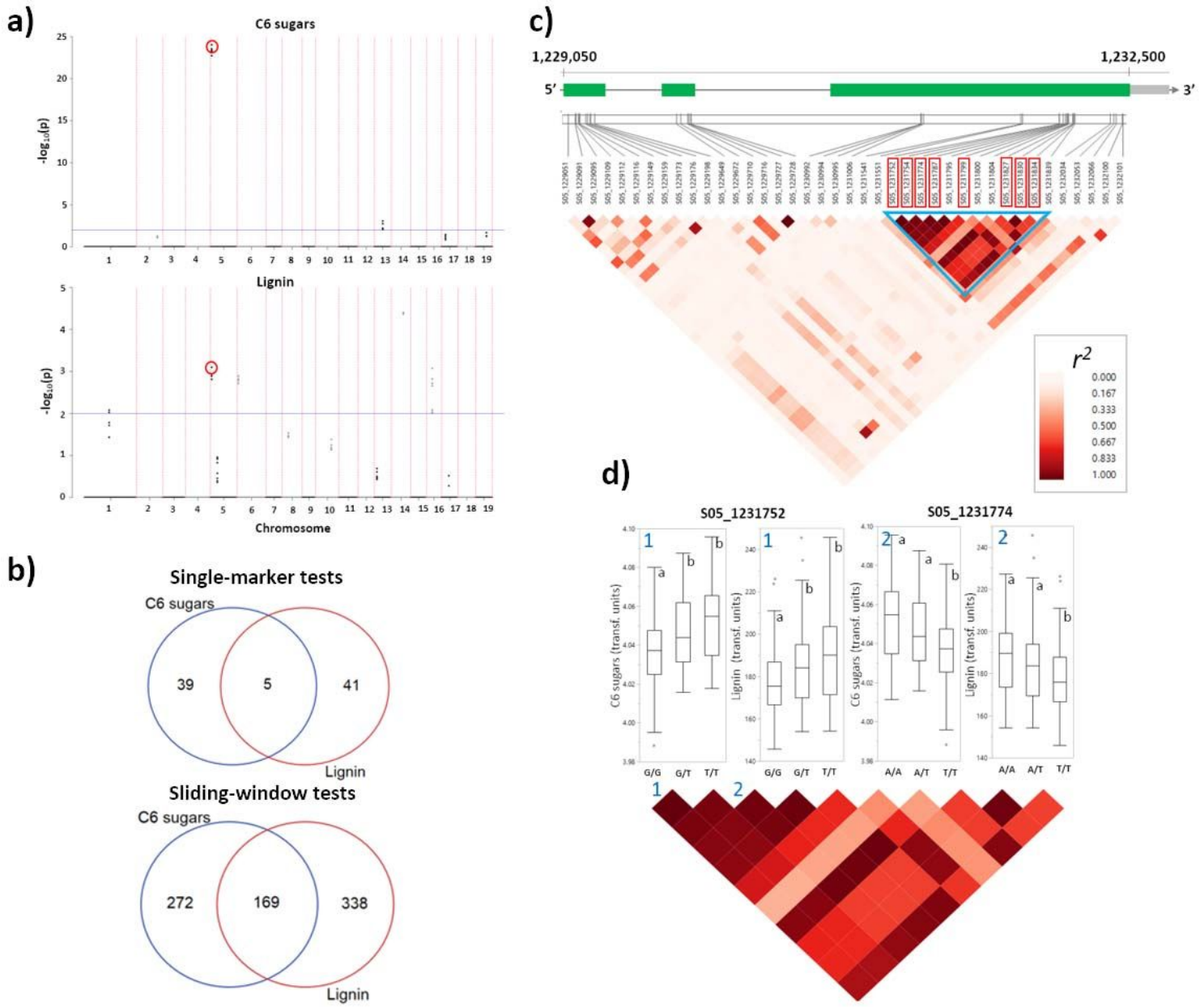

Figure 0

Analysis of SNPs significantly associated with two traits. Characterization of Leucine Rich Repeat (LRR 1)//Leucine Rich Repeat (LRR 8) gene (Potri.005G015700) associated with C6 sugars and lignin. a) Manhattan plots for C6 sugars and lignin highlighting (red circle) the common windows containing SNPs for this gene. The horizontal blue line indicates a -log10 (p-value) of 2 (equivalent to p-value = 0.01). b) Venn diagrams for the number of unique significant SNP-markers associated with both C6 sugars and lignin, identified in single-marker and sliding-windows tests, respectively. c) LD heat map for the analyzed SNPs located at gene. Red bars at the top correspond to SNPs identified as significantly associated with both C6 sugars and lignin by single-marker association tests. d) Detailed view for the blue light triangle depicted in c). Numbers 1 and 2 are the markers S05_1231752 and S05_1231774, respectively. Boxplots show the effects of genotypes on traits. Different letters indicate significant differences among adjusted means (Tukey's HSD test; α=0.001).

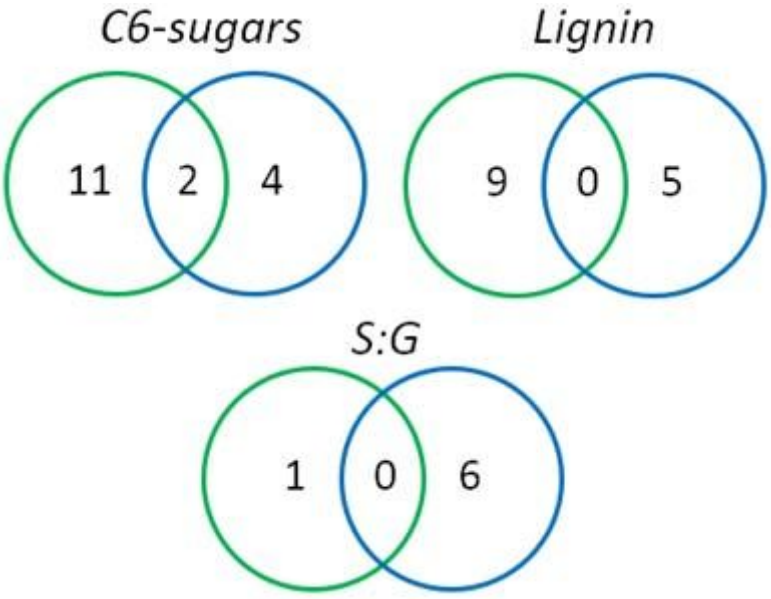

Figure 0

Venn diagrams for the comparison between the present study (blue circles) and the one carried out by Wegrzyn et al.[14] (green circles). Forty genes encoding enzymes involved in lignin and cellulose biosynthesis and cytoskeletal proteins were compared. Numbers in circles indicate the number of genes containing significant SNPs. A detailed list is given in Table S5.

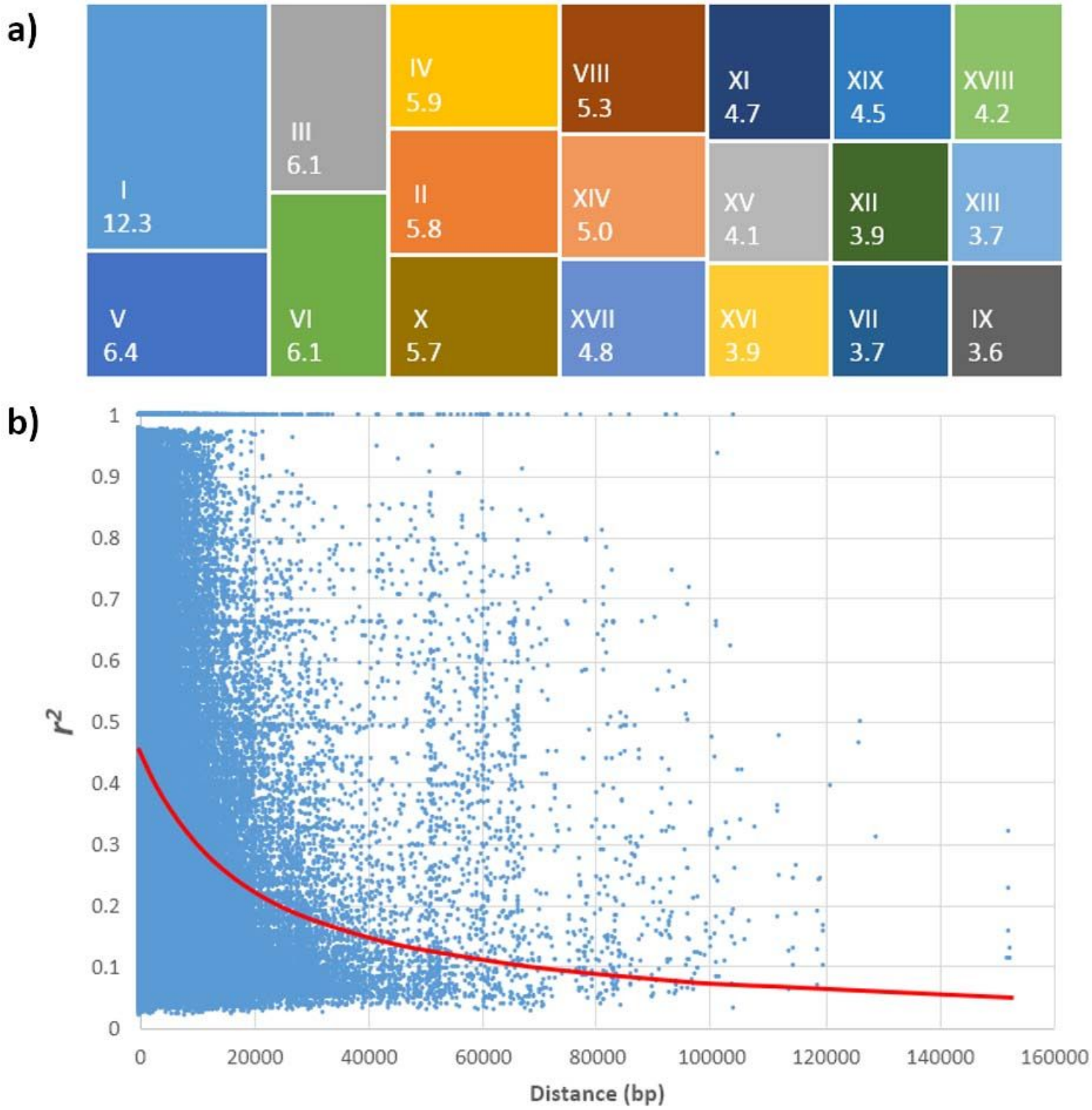

Figure 0

SNP genotyping and LD decay. a) Relative contribution (in percentage) of each chromosome to the total (813,280) of analyzed SNP-markers. b) Representative LD plot depicting the LD decay for Chromosome 12. The red line indicates the adjusted model for the significant correlations between SNP pairs.

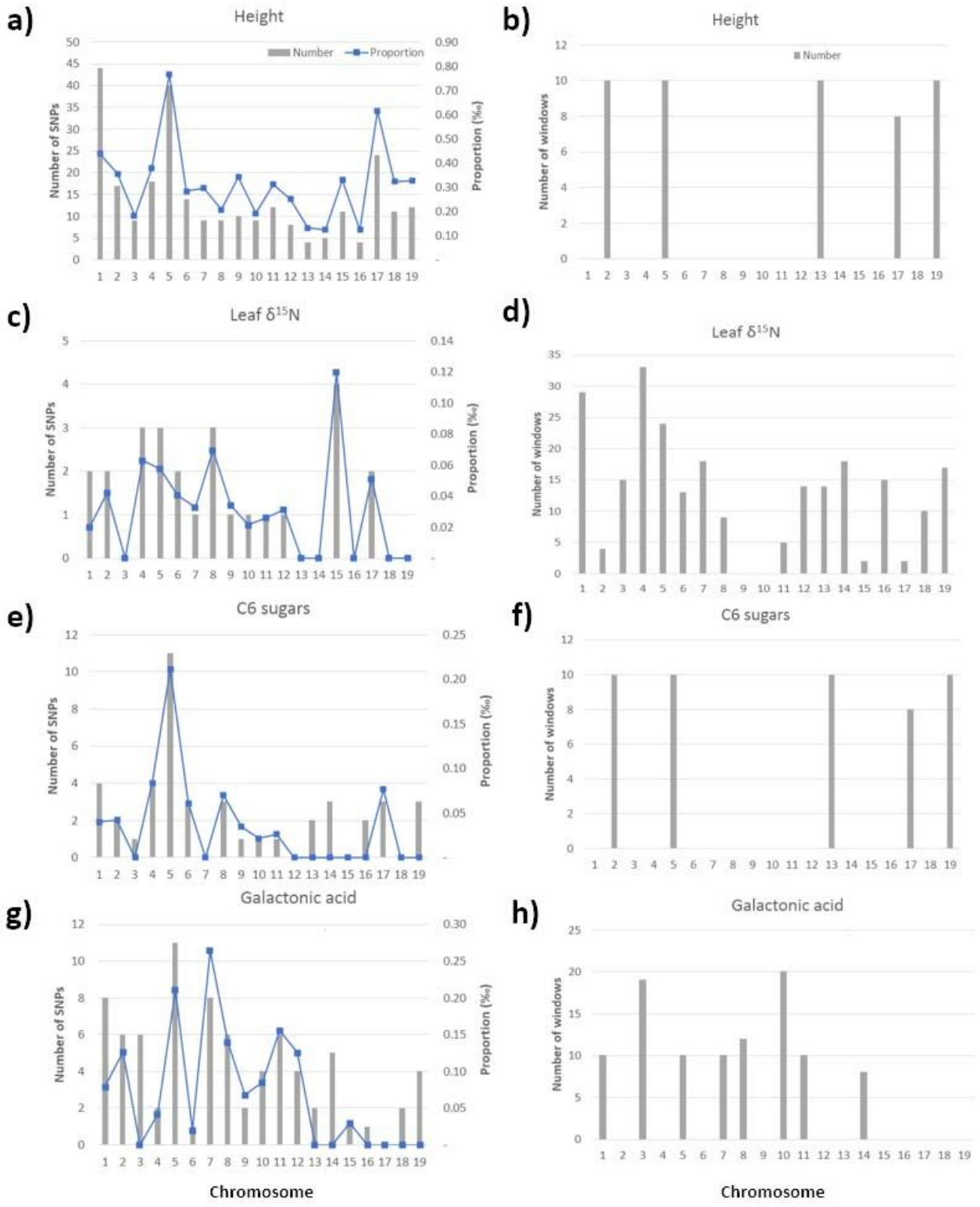

Figure 0

Number of significant single-SNPs (left) and sliding windows (right) associated with a selected set of traits for growth (a-b), stable isotopes parameters (c-d), chemical components of wood (e-f) and selected metabolites (g-h). Blue line at the left graphs indicates the proportion (‰) of significant SNP calculated on the total of analyzed SNP per chromosome. Significance thresholds considered a p-value < 0.0001

and q-value < 0.1, for single SNPs, and a q-value < 0.1 for sliding windows. Detailed information is provided in Tables S1 and S2.

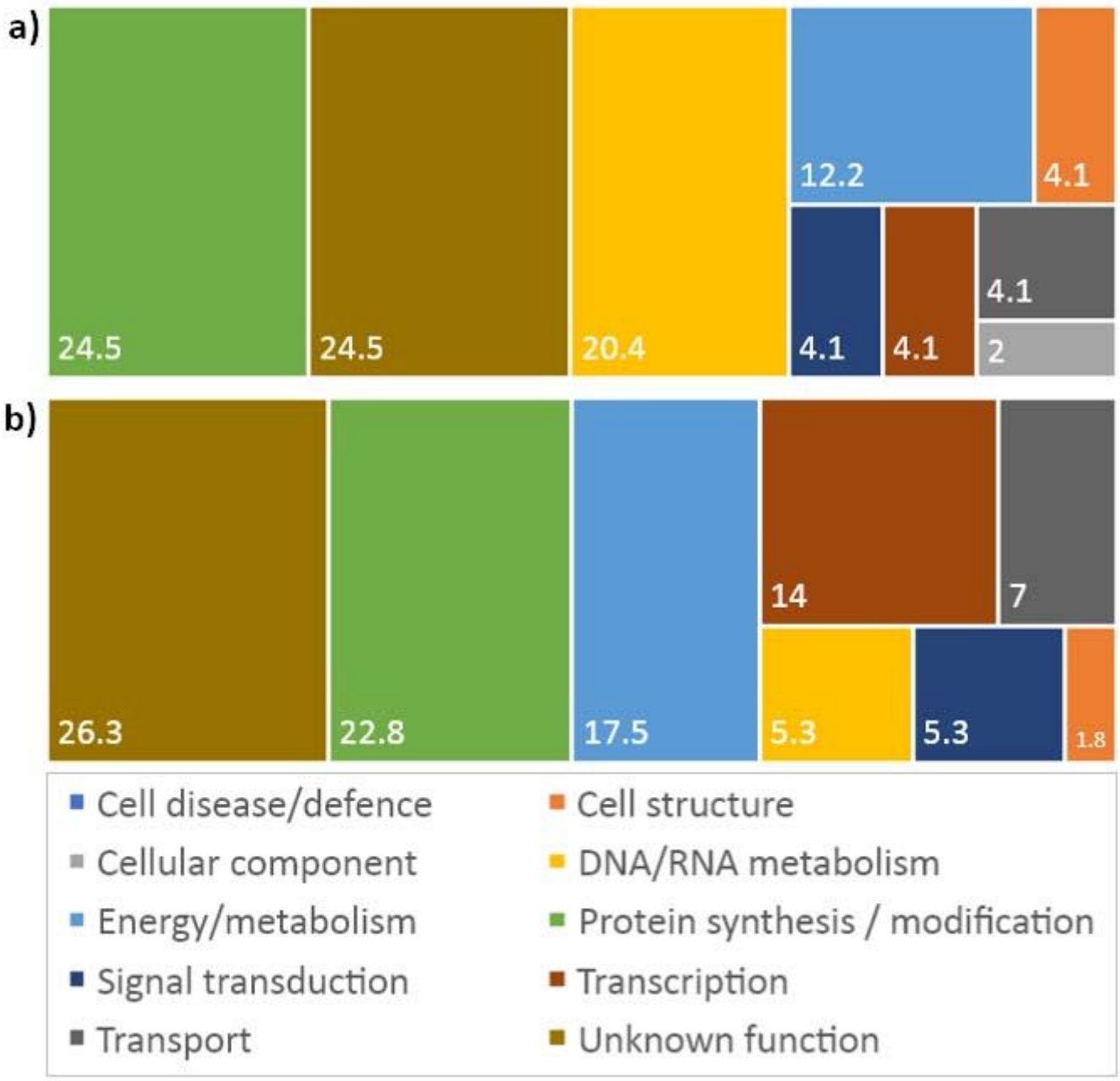

Figure 0

Main functional classes for the top three single-SNPs identified across all the analyzed phenotypes. a) Single SNP-marker associations. b) Sliding window analyses. Numbers represent percentages on total top three single-SNPs or sliding windows. Detailed information is provided in Tables S3 and S4.

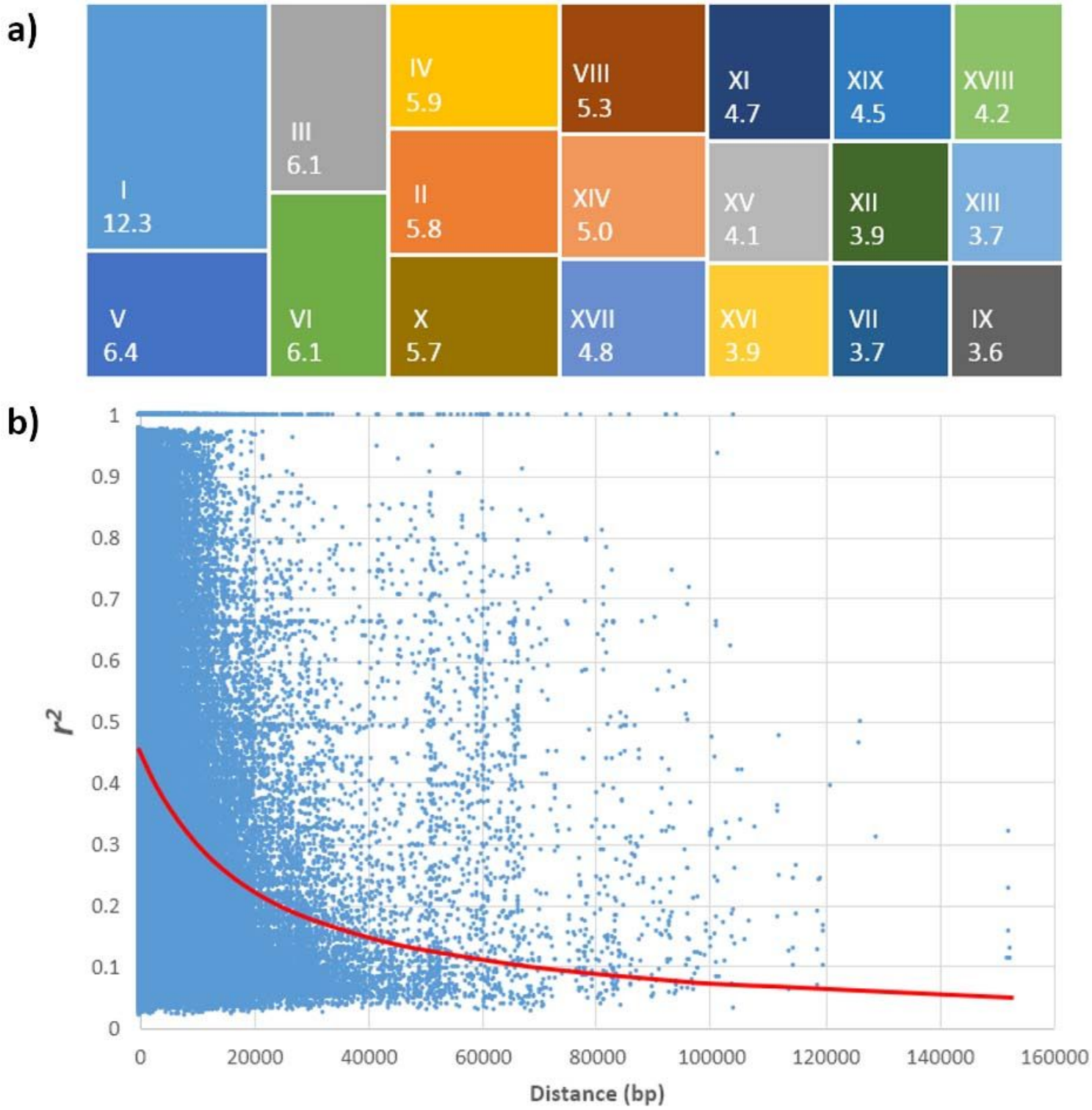

Figure 1

SNP genotyping and LD decay. a) Relative contribution (in percentage) of each chromosome to the total (813,280) of analyzed SNP-markers. b) Representative LD plot depicting the LD decay for Chromosome 12. The red line indicates the adjusted model for the significant correlations between SNP pairs.

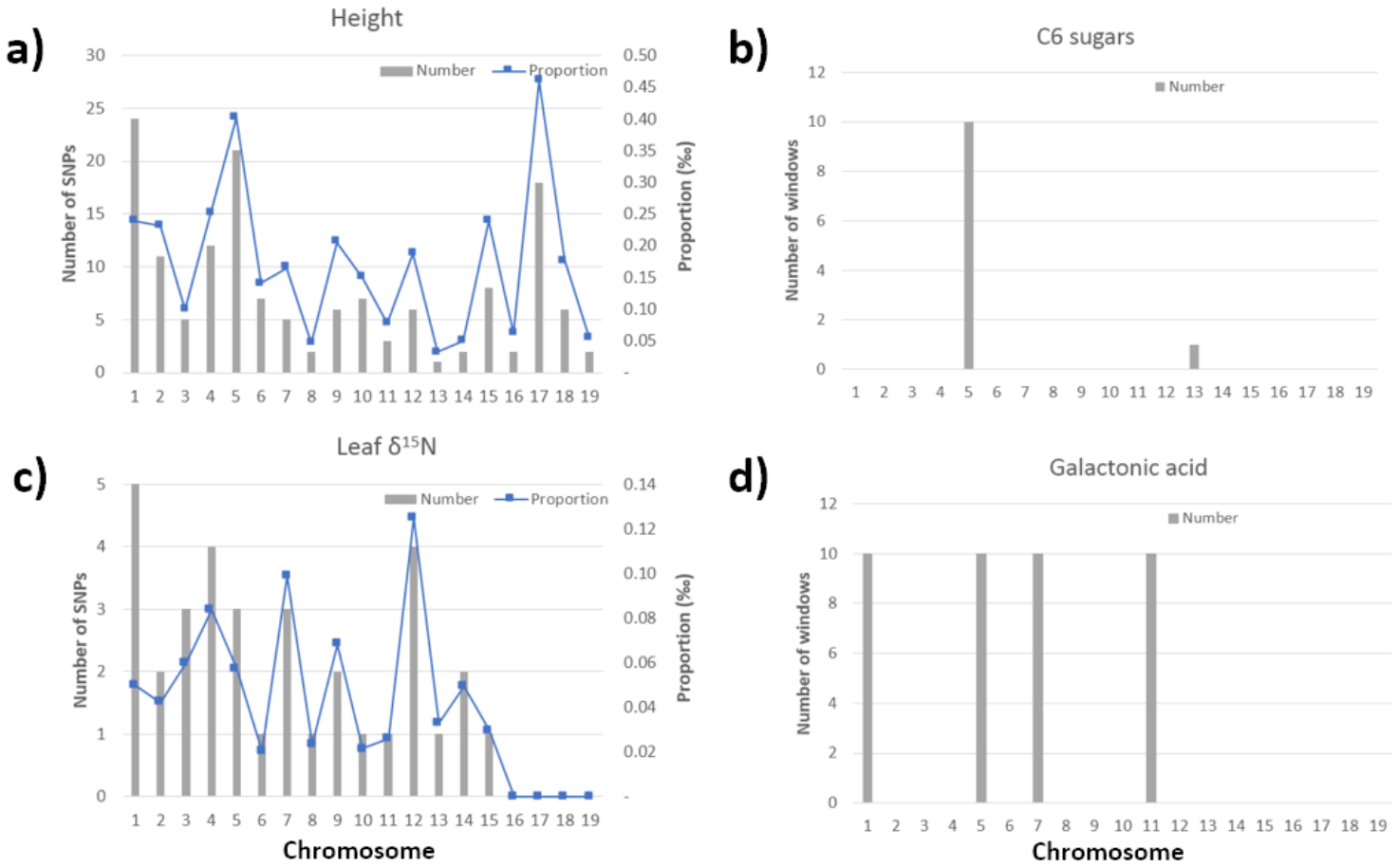

Figure 2

Number of significant single-SNPs (left) and sliding windows (right) associated with a selected set of traits for growth (a), stable isotopes parameters (c), chemical components of wood (b) and selected metabolites (g). Blue line at the left graphs indicates the proportion (‰) of significant SNP calculated on the total of analyzed SNP per chromosome. Significance thresholds considered a p-value < 6.1479E-8 for single-SNPs (a and c), and 1.04E-03 and 5.05E-04 for C6-sugars (b) and GAc (d) sliding windows, respectively. Detailed information is provided in Tables S1 and S2.

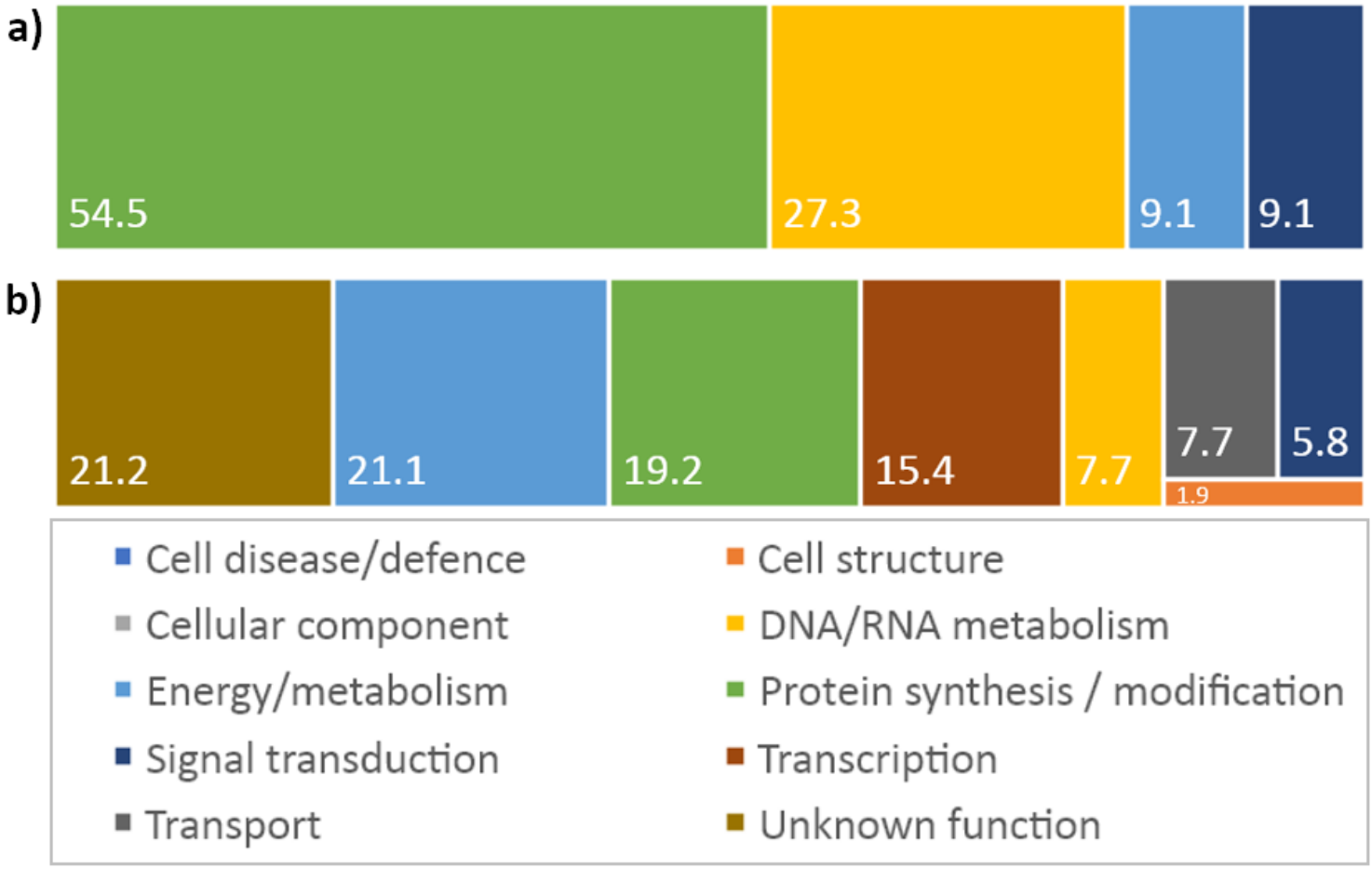

Figure 3

Main functional classes for the top three significant single-SNPs or sliding windows identified across all the analyzed phenotypes. a) Single SNP-marker associations. b) Sliding window analyses. Numbers represent percentages on total top three single-SNPs or sliding windows. Detailed information about specific SNP or windows is provided in Tables S3 and S4.

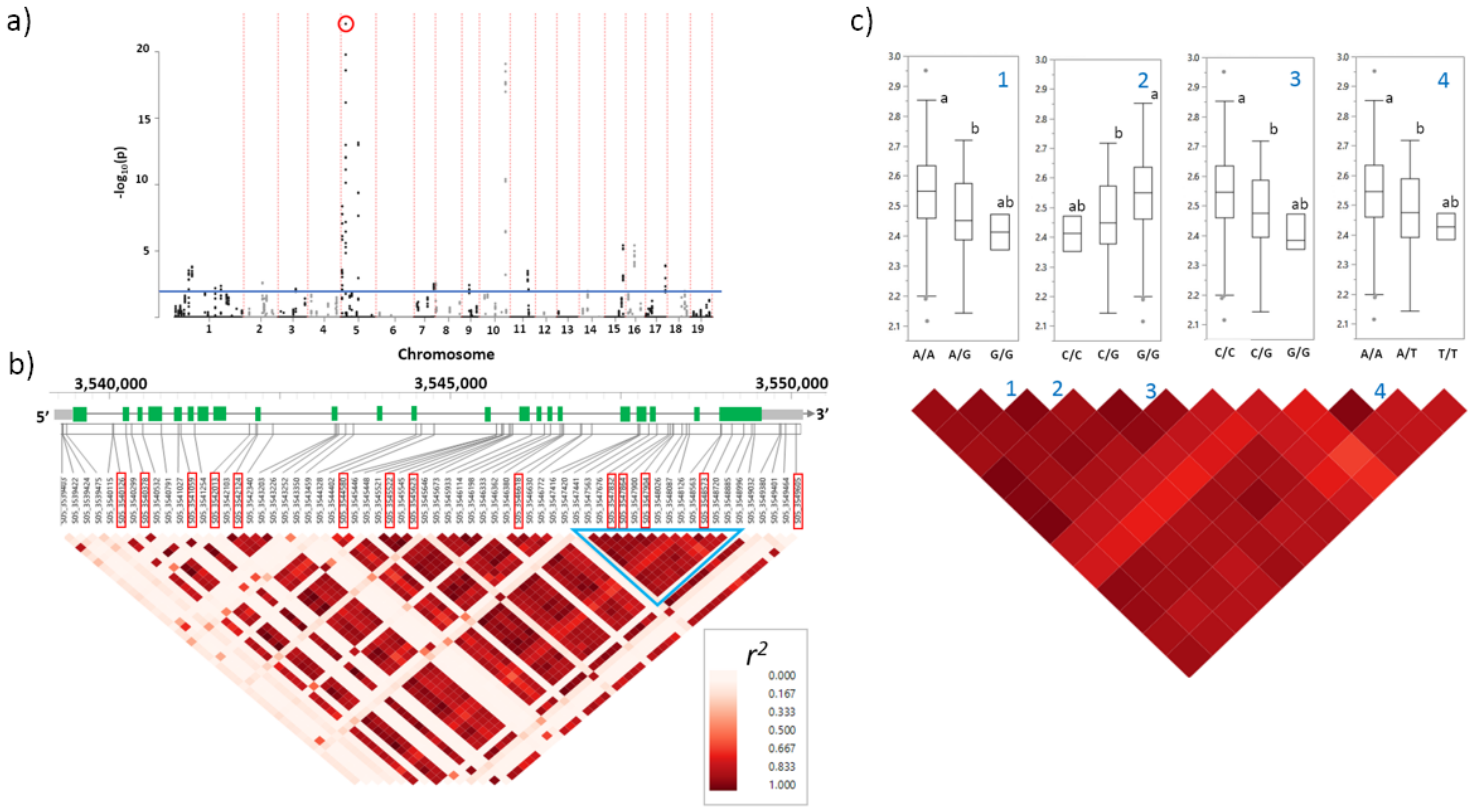

Figure 4

Detailed characterization of Similar to 5'-3' Exoribonuclease (XRN4) gene (Potri.005G048900) associated with leaf δ15N. a) Manhattan plot for leaf δ15N highlighting (red circle) the window containing significant SNPs for the gene. The horizontal blue line indicates a referential -log10 (p-value) of 2 (equivalent to p-value = 0.01). b) LD heat map for the analyzed SNPs located at gene. Red bars at the top correspond to SNPs identified as significantly associated with δ15N by single-marker association tests. c) Detailed view for the light blue triangle depicted in b). Numbers 1, 2, 3 and 4 are the markers S05_3547832, S05_3547864, S05_3547904 and S05_3548573, respectively. Boxplots shows the effects of genotypes on leaf δ15N. Different letters indicate significant differences among adjusted means (Tukey's HSD test; α=0.001).

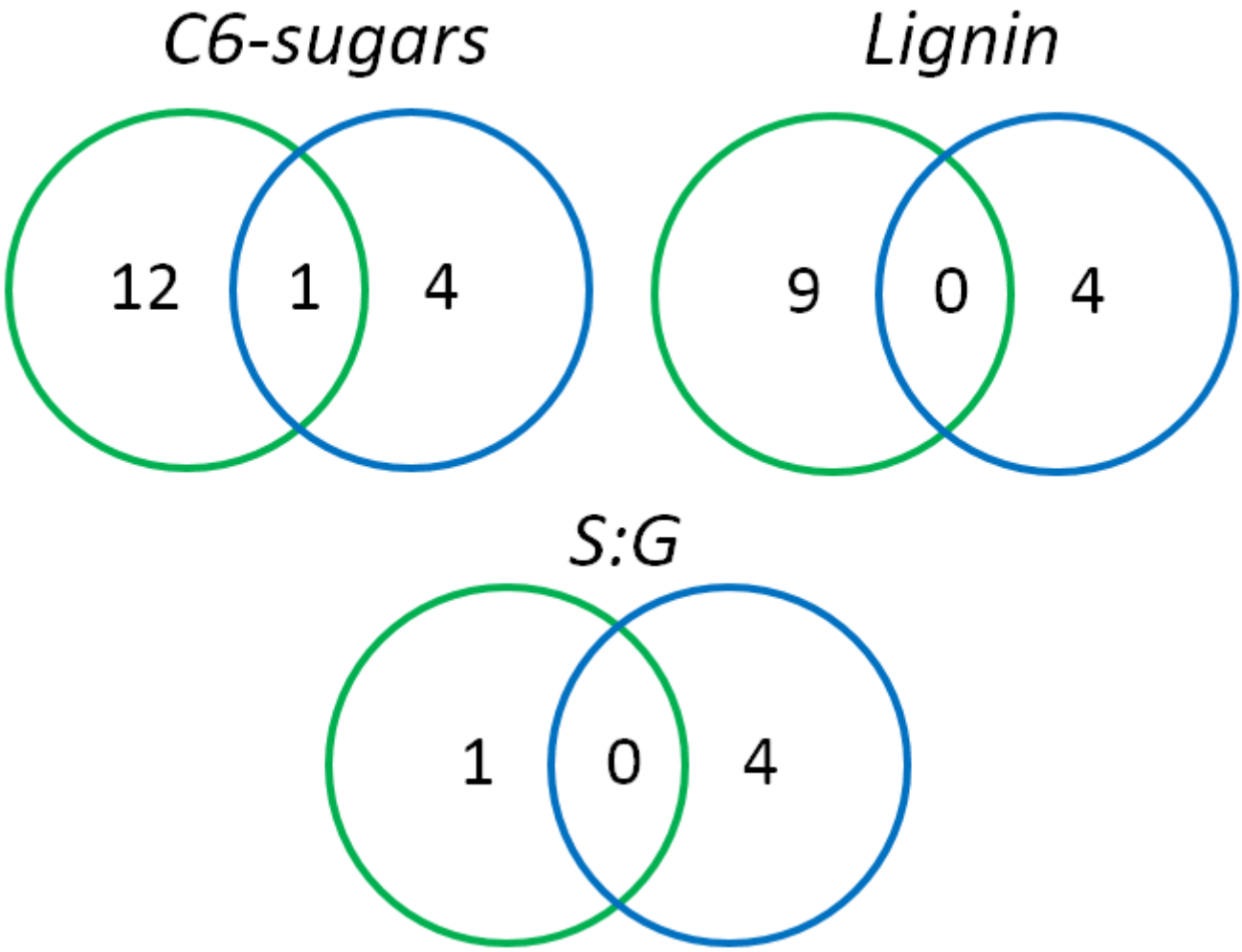

Figure 5

Venn diagrams for the comparison between the present study (blue circles) and the one carried out by Wegrzyn et al.[18] (green circles). Forty genes encoding enzymes involved in lignin and cellulose biosynthesis and cytoskeletal proteins were compared. Numbers in circles indicate the number of genes containing significant SNPs ($p<0.0001$). A detailed list is given in Table S5.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- FigS4.pdf
- FigS1.pdf
- FigS3.pdf
- FigS5.pdf

- TableS5.pdf
- FigS2.pdf
- Equation1.png
- supplement2.pdf
- supplement3.pdf
- supplement4.pdf
- supplement5.pdf
- supplement7.pdf
- supplement8.jpg
- supplement8.xlsx
- supplement9.jpg
- supplement9.pdf