Identification of genetic markers and wood properties that predict wood biorefinery potential in aspen bioenergy feedstock (*Populus tremula*).

Sacha Escamez[1,2], Mikko Luomaranta[1], Niklas Mähler[1], Madhavi Latha Gandla[3], Kathryn M. Robinson[1], Zakiya Yassin[4], Thomas Grahn[4], Gerhard Scheepers[4], Lars-Göran Stener[5], Leif J. Jönsson[3], Stefan Jansson[1], Nathaniel Street[1], Hannele Tuominen[1,2]*

Author Affiliations
1. Umeå Plant Science Centre (UPSC), Department of Plant Physiology, Umeå University, SE-901 87, Umeå, Sweden
2. Umeå Plant Science Centre (UPSC), Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 83, Umeå, Sweden
3. Department of Chemistry, Umeå University, SE-901 87, Umeå, Sweden
4. RISE AB, Drottning Kristinas väg 61 B, SE-114 28, Stockholm, Sweden
5. The Forestry Research Institute of Sweden, Ekebo, SE-268 90 Svalöv, Sweden
* Corresponding author: hannele.tuominen@slu.se

Abstract

Wood represents the majority of the biomass on lands, and it constitutes a renewable source of biofuels and other bioproducts. However, wood is recalcitrant to bioconversion, meaning that feedstocks must be improved. We investigated the properties of wood that affect bioconversion, as well as the underlying genetics, to help identify superior biorefinery tree feedstocks. We recorded as many as 65 wood-related and growth traits in a population of European aspen natural genotypes. These traits included three growth and field performance traits, 20 traits for wood chemical composition, 17 traits for wood anatomy and structure, and 25 wood saccharification traits as indicators of bioconversion potential. We used statistical modelling to determine which wood traits best predict bioconversion yield traits. This way, we identified a core set of wood properties that predict bioprocessing traits. Several of these predictor traits showed high broad-sense heritability, suggesting potential for genetic improvement of feedstocks. Finally, we performed genome-wide association study (GWAS) to identify genetic markers for yield traits or for wood traits that predict yield. GWAS revealed only a few genetic markers for saccharification yield traits, but many more SNPs were associated with wood chemical composition traits, including predictors traits for saccharification. Among them, 16 genetic markers associated specifically with lignin chemical composition were situated in and around two genes which had not previously been associated with lignin. Our approach allowed linking aspen wood bioprocessing yield to wood properties and the underlying genetics, including the discovery of two new potential regulator genes for wood chemical composition.

Key words:

Aspen (*Populus tremula*), biorefinery, genome-wide association study (GWAS), lignocellulosic biomass, saccharification, statistical modelling, wood properties.

Introduction

Lignocellulose of vascular plants, mainly in the form of wood, represents the majority of the biomass on land (Bar-On et al., 2018). This biomass reservoir contains mostly three types of natural polymers: cellulose, hemicelluloses and lignin, each of which can be converted into precursors for biofuels and other bioproducts (Percival Zhang, 2013). However, the processes for deconstructing these polymers into usable units remain costly due to structural and chemical hindrance, a problem known as biomass recalcitrance (McCann and Carpita, 2015).

Overcoming biomass recalcitrance requires the identification of less recalcitrant feedstocks as well as knowledge on the biological basis of lignocellulose recalcitrance (Van Acker et al., 2014, Wilkerson et al., 2014, Escamez et al., 2017, Meng et al., 2017, Yoo et al., 2017, Wang et al., 2020). Fast growing trees from the *Populus* genus (poplars, aspens, and hybrids) represent promising feedstocks (Mola-Yudego et al., 2017) on account of their lignocellulose composition (Sannigrahi et al., 2010) and of their advanced domestication and cultivation techniques (Dickmann, 2006). Furthermore, the genomes of numerous *Populus* species have been sequenced (Tuskan et al., 2006, Ma et al., 2013, Wullschleger et al., 2013, Yang et al., 2017, Wang et al., 2018a, Lin et al., 2018, Liu et al., 2019, Qiu et al., 2019, Hou et al., 2020, Zhang et al., 2020), enabling investigation of the genetics underlying lignocellulose properties and recalcitrance.

Our knowledge of the genetic basis for plant traits has greatly advanced thanks to genome-wide association studies (GWAS), which relate variation in traits to variation in the sequence of the genomes of different individuals, at a single nucleotide resolution (Nordborg and Weigel, 2008, Tuskan et al., 2019). These variations of nucleotide composition at single loci between the compared individuals, also known as single nucleotide polymorphisms (SNPs), can represent genetic markers for quantitative variation in traits, or even reveal involvement of genes into shaping a quantitative trait (Nordborg and Weigel, 2008, Tuskan et al., 2019).

In a striking example, GWAS of the timing of budset identified a single locus explaining the majority of local adaptation along a latitudinal gradient in a Swedish population of European aspen *Populus tremula* (Wang et al., 2018a). However, individual loci found by GWAS usually explain only a fraction of the total trait variance, and there is often a large portion of the genetically heritable variance that remains undetermined by significant associations (Nordborg and Weigel, 2008, Du et al., 2018). Nevertheless, finding SNPs associated with only a fraction of the variation in traits of interest could still lead to progress through marker assisted selection (MAS) or genomics assisted selection (GAS) for beneficial wood properties (Du et al., 2018).

In *Populus trichocarpa,* GWAS revealed SNPs and genes significantly associated with four wood chemical composition traits (Guerra et al., 2019). Still in *Populus trichocarpa*, focusing on a limited part of the genome consisting of genes expressed in wood, associations were discovered between SNPs and 16 wood chemical composition and wood structure traits (Porth et al., 2013). Four wood chemical composition traits were also linked to SNPs by GWAS in *Populus nigra* (Guerra et al., 2013) and *Populus deltoides* (Fahrenkrog et al., 2017). Xie et al. (2018) re-evaluated previous associations in *Populus trichocarpa* (Porth et al., 2013, Muchero et al., 2015) by focusing on a chromosome known to harbour quantitative trait loci (QTL) for lignin composition, resulting in the identification and characterization of a new transcriptional regulator of lignin biosynthesis. Using multivariate GWAS, whereby traits can be aggregated into multi-traits for GWAS (Porter and O'Reilly, 2017, Chhetri et al., 2019), 19 SNPs related to 13 genes were identified in association to wood anatomical properties of a *Populus trichocarpa* natural population (Chhetri et al., 2020).

Advances in genome (re)sequencing and statistical methods for finding associations in GWAS have greatly facilitated these recent findings (Du et al., 2018, Lin et al., 2018). Yet, the emerging picture of the genetics underlying wood properties and bioconversion potential of *Populus* remains limited, in

parts due to our limited capacity to generate precise quantifications of specific traits for entire tree populations (Du et al., 2018, Tuskan et al., 2019), a problem known as the "phenotyping bottleneck" (Furbank and Tester, 2011). For example, lignin is composed of different types of monomers which polymerize together into a heteropolymer (Boerjan et al., 2003). Hence, measuring the total amount of lignin in wood represents a coarse measure, whereby finer traits such the abundance of the different types of lignin monomers remain hidden (Tuskan et al., 2019). More numerous and more specific traits more likely allow identification of significant associations in GWAS (Du et al., 2018, Tuskan et al., 2019). More precise phenotyping also allows better characterization of the relationships between traits, for example to identify which wood chemical composition and structure traits determine wood bioconversion potential (Escamez et al., 2017).

Here, we present large-scale phenotyping efforts, monitoring as many as 65 traits related to wood properties, tree growth, and wood saccharification in a collection of natural European aspen genotypes collected across Sweden. Through multivariate analyses and mathematical modelling, we identified wood chemical composition and structural traits predictive of recalcitrance as well as whole stem bioconversion potential. Through GWAS, we identified genetic loci linked to wood properties predictive of bioconversion, including in genes not previously linked to lignocellulose.

Materials and Methods

Plant material

The Swedish Aspen (SwAsp) collection consists of previously gathered *Populus tremula* aspen natural genotypes from 12 locations across Sweden (Luquez et al., 2008). These aspen genotypes had been clonally propagated from root cuttings, and then grown in a randomized block experiment in two plantations in southern (Ekebo, 55.9 °N) and northern (Sävar, 63.4 °N) Sweden, with originally at least four to five biological replicates per genotype, of which three to five were successfully established (Luquez et al., 2008, Wang et al., 2018a).

After ten years of growth in the Ekebo garden, tree height and diameter at breast height (DBH) were measured (Dataset S1), and wood samples were collected (Fig. S1). 79 cm above ground, a 1 cm thick section of the stem was collected, and the south-western facing quarter of the stem section was aliquoted for wood chemical composition analyses. In addition, 90 cm above ground, another piece of stem was harvested for analysis of wood anatomical and structural properties.

Wood chemical composition analyses

The wood quarters selected for compositional analyses were manually de-barked and cut into roughly match-size wood pieces and freeze-dried (CoolSafe Pro 110-4, LaboGene A/S, Denmark). These wood pieces were homogenized by coarse milling (Retsch ZM200 centrifugal mill, Retsch GmbH, Germany), and sieved (Retsch AS200) into two different particle size fractions. The fraction of particle size between 0.1 mm and 0.5 mm was aliquoted for subsequent saccharification experiments (see below), while the fraction of particle size under 0.1 mm was aliquoted for pyrolysis coupled with gas chromatography followed by mass spectrometry analysis (pyrolysis-GC/MS) and monosaccharide composition analysis.

Total carbohydrate content, lignin content, lignin composition, and content of other phenolics were determined by pyrolysis coupled with gas chromatography followed by mass spectrometry analysis (pyrolysis-GC/MS) as previously described (Gerber et al., 2016). Briefly, 40 µg - 80 µg of fine wood homogenized powder was loaded into an autosampler (PY-2020iD and AS-1020E, Frontier Labs, Japan), allowing a sub-sample (~1 µg) into the pyrolizer of the GC/MS apparatus (Agilent,

7890A/5975C, Agilent Technologies AB, Sweden). Following pyrolysis, the samples were separated along a DB-5MS capillary column (30 m × 0.25 mm i.d., 0.25-µm-film thickness; J&W, Agilent Technologies), and scanned by the mass spectrometer along the m/z range 35 – 250. The GC/MS data was processed as previously described (Gerber et al., 2012). To make the samples comparable with one another, each peak's area was normalized to the total peak area considering all peaks, set as 100%, in each sample.

Cell wall monosaccharides were quantified following the trimethylsilyl (TMS) derivatization method as described previously (Gandla et al., 2015). Briefly, fine wood powder was washed with HEPES buffer (4 mM, pH 7.5) containing 80% ethanol, as well as methanol:chloroform 1:1 (v:v) and acetone to generate alcohol insoluble residues (AIRs), which were then dried. To avoid contamination with glucose from potential starch reserves, the AIRs were treated with 1 unit per AIR mg of type I α-amylase (Roche 10102814001, Roche GmbH, Germany). The de-starched AIRs were methanolysed using 2 M HCl/MeOH at 85 °C for 24 h, and inositol was also methanolysed to serve as internal standard. Following repeated washes with methanol, the AIRs and inositol standards were silylated using Tri-sil reagent (3-3039, SUPELCO, Sigma-Aldrich GmbH, Germany) at 80 °C for 20 min. The solvent was evaporated under a stream of nitrogen and pellets were dissolved in 1 ml hexane and filtered through glass wool. The filtrates were evaporated until only 200 µl remained, of which 0.5 µl were analysed by GC/MS (7890A/5975C; Agilent Technologies AB, Sweden) according to Sweeley et al. (1966).

Saccharification assays

Saccharification assays without or with acid pretreatment of the woody biomass were performed following a previously established methodology (Gandla et al., 2015). In short, 50 mg of dry wood powder (moisture analysis performed using an HG63 moisture analyser, Mettler-Toledo, USA) with particle size between 0.1 mm and 0.5 mm were pretreated with 1% (w/w) sulphuric acid during 10 min at 165 °C in a single-mode microwave system (Initiator Exp, Biotage, Sweden), or remained untreated. The pretreated samples were centrifuged to allow separation of the solid fraction from the pretreatment liquid. The solid fraction was washed with ultrapure water and sodium citrate buffer (50 mM, pH 5.2). The washed pretreated solid fraction as well as the untreated samples were enzymatically hydrolysed 72 h at 45 °C under agitation, using 25 mg of a 1:1 (w/w) mixture of the liquid enzyme preparations Celluclast 1.5 L (measured CMCase activity of 480 units per gram of liquid enzyme preparation, following Ghose, 1987) and Novozym 188 (measured β-glucosidase activity of 15 units per gram liquid enzyme preparation, following Mielenz (2009)) (Sigma-Aldrich). Sodium citrate buffer (50 mM, pH 5.5) was added to reach 1 g of final reaction mixture. During enzymatic saccharification, samples were collected at 2 h and 72 h. Glucose production rates were determined at 2 h using an Accu-Chek ®Aviva glucometer (Roche Diagnostics Scandinavia AB, Sweden) by calibration with different concentrations of glucose standard solution. Monosaccharide (arabinose, galactose, glucose, xylose and mannose) yields in pretreatment liquids and enzymatic hydrolysates collected at 72 h were determined using a high-performance anion-exchange chromatography (HPAEC) system equipped with pulsed amperometric detection (Ion Chromatography System ICS-5000, Dionex, USA) according to a previously described procedure (Wang et al., 2018b)

Anatomical and structural characterisation

As previously described (Lundqvist et al., 2010, Escamez et al., 2017), anatomical and structural features were determined on parallelepipedal wood pieces across the stem diameter using SilviScan (CSIRO, Australia). The wood sample strips were mounted on computer controlled motorised stages

and scanned for information on wood property variations along a stem radius. Characterisation was performed on three separate units representing different measurement methods: (i) a cell scanner with a video microscope for measurement of the numbers and sizes of fibres and vessels, (ii) a density scanner recording X-ray absorption images for measuring wood density, and (iii) a diffraction scanner recording X-ray diffraction images for measuring the microfibril angle.

Statistical estimations of genetic parameters

The genetic parameters for each trait were estimated statistically based on measurements on individual trees for each genotype according to the model:

$Y_{ijk} = \mu + b_i + c_j + e_{ijk}$ (Equation 1)

Where Y_{ijk} is the observation k in block i for clone j, μ is the mean of the trait in this trial, b_i is the fixed effect of block i, c_j is the random effect of clone j (normally and independently distributed with mean 0 and variance V_c; NID[0,V_c]), and e_{ijk} is the random error term for observation ijk (NID[0,V_e]). The variances V_c and V_e were estimated for each trait according to the Restricted Maximum Likelihood (REML) method using the ASREML software (Gilmour et al., 1997). To estimate genetic parameters, we considered $V_c = V_G$ (the genotypic variance among clones for the trait) and $V_e = V_E$ (the environmental variance for the trait).

For each trait, broad-sense heritability (H^2) was estimated by dividing genotypic variance (V_G), by the total variance of this trait V_T; where $V_T = V_G + V_E$:

$H^2 = V_G/V_T$ (Equation 2)

The genotypic coefficient of variation (CV_G) for a trait was calculated by dividing the trait's genotypic standard deviation $\sqrt{V_G}$ by the mean value of the trait's measurements ($\bar{x}$), and multiplying the result by 100:

$CV_G = \sqrt{V_G} \cdot 100/\bar{x}$ (Equation 3)

The genetic correlation (r_G) between trait 1 with genotypic variance V_{G1} and trait 2 with genotypic variance V_{G2} was calculated by diving the genotypic genetic covariance (cov_{G1G2}) between these traits by the square root of the product of the individual genetic variances of these traits:

$r_G = cov_{G1G2}/\sqrt{(V_{G1} \cdot V_{G2})}$ (Equation 4)

Statistical comparisons, multivariate analyses and statistical modelling

Multiple pairwise comparisons were carried by so called protected post-ANOVA Fisher's LSD tests, whereby Fisher's LSD tests for multiple comparisons are only performed if the ANOVA returns $p < 0.05$, using Minitab 17 (Cleverbridge AG, Germany). This method has better risk mitigation for both type I and type II errors, especially when having between three and ten biological replicates, than other commonly used tests (Carmer and Swanson, 1973).

Multivariate analyses using all wood traits to predict glucose release by saccharification, or the estimated total-wood glucose yield from an entire tree trunk (TWG, Escamez et al., 2017), were performed using Orthogonal Projections to Latent Structures (OPLS) regression (Trygg and Wold,

2002), as previously described for a different tree population (Escamez et al., 2017), with 1 + 3 components.

For predicting saccharification traits, or TWG, from only a subset of traits (Dataset S2), we followed a previously established methodology (Escamez et al., 2017). In short, using the R statistical software, we attempted predicting each of the selected yield traits (Dataset S2) with at least 30 different models of three different types, relying on subsets of wood chemical, anatomical and structural traits. The three types of models consisted in multiple linear regressions, as well as the machine learning algorithms random forests (Breiman, 2001) and generalized additive models (GAMs, Hastie and Tibshirani, 1986).

The >90 different models for each trait of interest were assessed for their predictivity by leave-one-out cross-validation, whereby values for a tree genotype were removed, the model was generated from all the remaining tree genotypes, and finally the value from the left-out genotype was predicted. Repeating this operation until every single tree genotype had been left out once allowed measuring the prediction error across the entire dataset. The ratio of prediction error to the total variation in the dataset was then expressed as the Q2 value, which can range between 1 (perfect predictions) and minus infinity, with Q2 > 0.5 being conventionally considered as the threshold for significantly acceptable prediction accuracy. The most predictive models of each type are presented in Dataset S2, for each trait of interest.

GWAS

Previous whole genome sequencing of 104 SwAsp clones, followed by high quality re-sequencing of these trees, resulted in 94 unrelated individual genotype sequences for GWAS (Mähler et al., 2017, Grimberg et al., 2018, Wang et al., 2018a, Mähler et al., 2020). The single nucleotide polymorphism (SNP) calling was performed as previously described (Wang et al., 2018a). This procedure yielded a total of 4,425,109 bi-allelic SNPs with minor allele frequency > 5% (Wang et al., 2018a), annotated by intersection with browser extensible data (BED) file from the genome annotation of the *Populus tremula* reference genome (Lin et al., 2018) (available at http://popgenie.org; Sjödin et al., (2009)). For the purpose of annotations, SNPs were considered as intergenic if they laid further than 2 kbp away from a gene, while SNPs within 2 kbp of a gene were considered associated with that gene.

Correlations between SNPs and phenotypes were estimated by genome-wide association study (GWAS). GWAS was performed by considering each of the median of each measured trait for each genotype as the dependent variable in a linear mixed model regression, using GEMMA (Zhou and Stephens, 2012, Zhou and Stephens, 2014). Relatedness among individuals, although weak, was accounted for as a covariate in the linear mixed model, as previously described (Wang et al., 2018a). In addition, latitude was also included as a covariate as previously described (Wang et al., 2018a). False discovery rate (FDR) of each association was calculated as the "q-value" using R (Storey et al., 2021) following the principle of the Benjamini-Hochberg procedure (Storey and Tibshirani, 2003). The effect size of each SNP (named "beta" in Dataset S3), for each trait, was also estimated as previously described (Wang et al., 2018a). Dataset S3 presents the 1000 SNPs with the lowest q-value, regardless of any arbitrary significance threshold, for each trait.

Gene expression analyses

A 2 cm piece from the base of the stem was collected from SwAsp clones grown *in vitro* in triplicates for one month. These stem pieces were immediately flash frozen and later ground in liquid nitrogen with mortar and pestle. From the frozen, ground stem samples, mRNA were extracted using the Qiagen RNeasy Plant Kit (Qiagen GmbH, Germany). Genomic DNA was removed using the kit's optional "on column" DNase treatment, followed by a second DNase treatment as a first step in the subsequent cDNA synthesis procedure of the QuantiTect Reverse Transcription Kit (Qiagen).

qPCR reactions were run using LightCycler 480 SYBR Green 1 Master (Roche) according to the manufacturer's instructions. Samples were loaded in a 96-well qPCR plate (Roche) as technical duplicates. The qPCR reactions were run and scored using a LightCycler 480 thermal cycler (Roche) over 45 cycles (initial denaturation at 98°C for 3 min, cycling: denaturation at 95°C for 5 s, annealing at 55°C for 10 s, elongation at 72°C for 30 s; melting curve from 45°C to 95°C with acquisitions every 1°C). To detect the transcript of the *Populus tremula* gene Potra000716g05617 (*Potra-pTAC13*), we used the following (5'-3') primers: forward-CTGGCCCTCTTGTGAGTAGC, reverse-CACAGTTGCCTTCCCAGTTT. To normalize the detected transcript amounts to reference genes (Livak and Schmittgen, 2001), we chose the ubiquitin biosynthetic gene Potra168132g27340 and the ribosomal protein 50s encoding gene Potra001573g13026, with the following primers: 50S_forward-CAAAGCCTTCAAAGCCCAAG, 50S_reverse-GCACTTACGAAGACGCAATG, ubiquitin-forward-GTTGATTTTTGCTGGGAAGC, ubiquitin-reverse-GATCTTGGCCTTCACGTTGT. The transcript levels of both reference genes were combined by geometric mean to increase robustness of this internal control (Vandesompele et al., 2002).

Results

Natural variation in 65 aspen stem growth, wood, and biorefinery traits. The 113 SwAsp genotypes were grown in replicates (originally, at least four clones per genotype) outdoors, in randomized blocks (Luquez et al., 2008). After ten years, we measured their stem height and diameter, and we analysed their wood's chemical composition (20 traits), wood structural and anatomical properties (17 traits), as well as the recovery of monosaccharides from wood saccharification without or after harsh acidic pretreatment (25 traits), amounting to a set of 64 traits (Fig. 1a; Dataset S1). Finally, we estimated total-wood glucose yield (TWG; Escamez et al., 2017; Fig.1a; Dataset S1). While glucose release provides information about biomass recalcitrance to saccharification, TWG provides a proxy for overall tree performance (Escamez et al., 2017).

All 65 traits showed variation between genotypes (Fig. 1a; Dataset S1), allowing to investigate both the correlations between different wood properties as well as the genetic causes for their variation. The variation in traits could not be explained by the geographical origin of the aspen genotypes (Fig. 1b), thus ruling out potential bias due to original sampling (when the collection was assembled, Luquez et al., 2008).

Prediction of yield from wood traits. To better characterize the basis for wood recalcitrance to bioprocessing, and the wood properties underlying yield, we performed multivariate analyses and statistical modelling of glucose release from saccharification, as well as TWG.

First, we employed orthogonal projections to latent structures (OPLS; Trygg and Wold, 2002) that consider all traits simultaneously, to get an overview of the relationships between wood properties and glucose release or TWG (Fig. S2a,b). A high proportion of the variation in glucose release (Fig. S2a) or TWG (Fig. S2b) could be explained (R2 = 0,56 and 0,52, respectively). However, leave-one-out cross validation revealed that neither glucose release nor TWG could be accurately predicted (Q2 = 0,17 and 0,29, respectively) when using the entire set of wood chemical composition, anatomy and structure traits. This result suggested that, rather than the entire set of wood traits, only a subset of wood traits was responsible for most of the variation in glucose release and TWG.

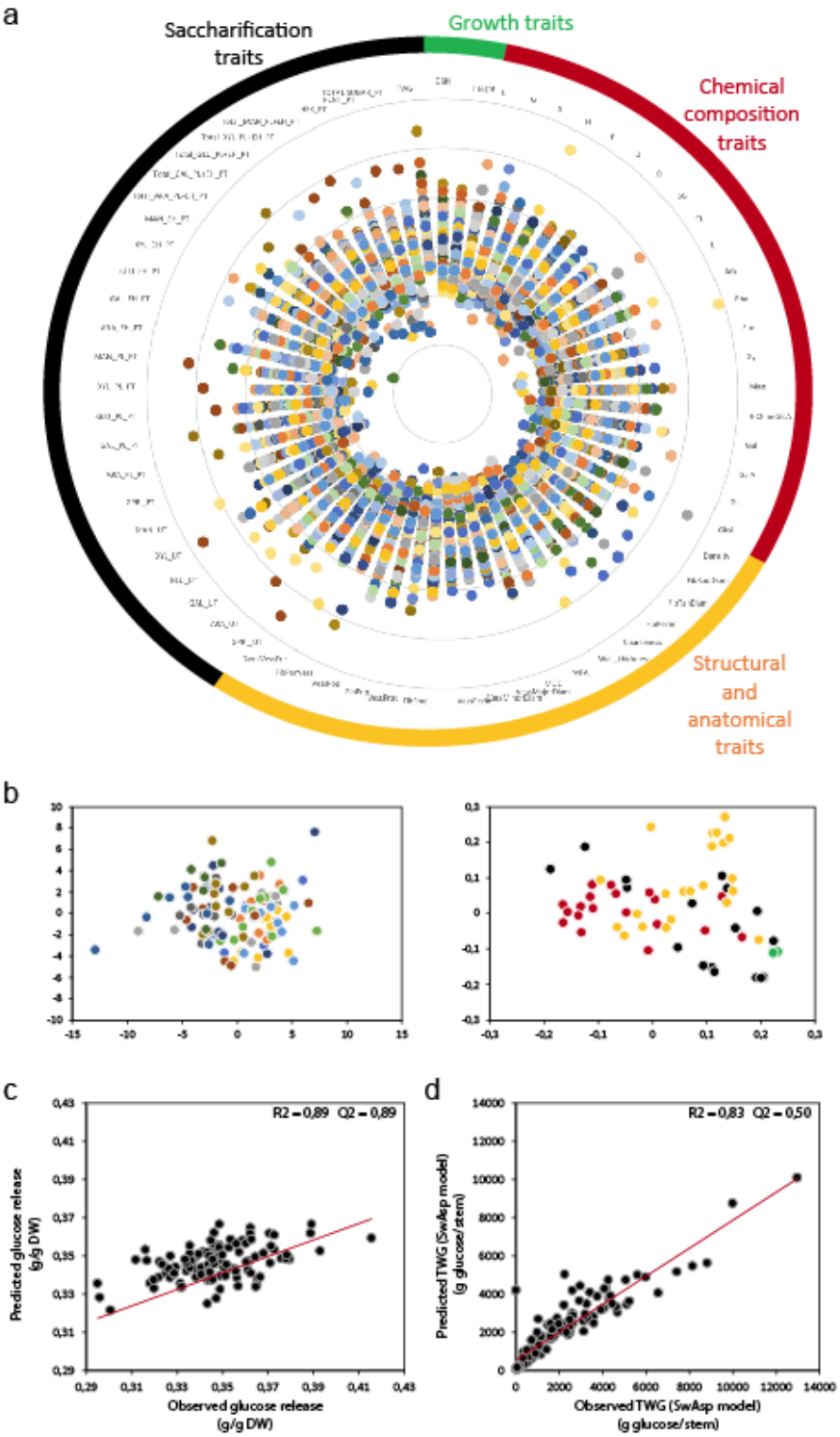

Figure 1. The SwAsp natural variants display a range of wood properties and saccharification yield independently of population structure.
(a) Measurements of 65 traits related to tree growth, wood chemical composition, wood anatomy and structure, and sugar yield from saccharification of woody biomass from 113 natural aspen variants collected across Sweden and grown for 10 years in a common garden. Each dot represents the median scaled and centered measurement of a trait for one genotype (z-transformation across the tree population for each trait), revealing wide variation for each trait. Colored labels around the plot indicate categories of traits (chemical composition, structure and anatomy, saccharification and growth), showing that all categories displayed variation to similar extents.
(b) Principal Component Analysis (PCA) showing that the SwAsp genotypes differ from each other (left) based on their wood properties and saccharification (right). Colors on the PCA scatter plot (left) indicate location of origin (all trees come from one of 12 locations in Sweden). Dots on the coefficients scatter plot (right) indicate traits, while their colors indicate which trait category they belong to (as in (a)).
(c) Scatter plot showing the correlation between the observed glucose release after pretreatment for the SwAsp trees (x-axis), and the corresponding prediction (y-axis) based on wood properties using a linear statistical model. R2 indicates the variance explained, while Q2 reflects the predictive accuracy of the model from leave-one-out cross validation.
(d) Scatter plot showing the correlation between the observed TWG for the SwAsp trees (x-axis), and the predicted TWG (y-axis) based on wood properties using a newly developed statistical model (SwAsp model; this study). R2 indicates the variance explained, while Q2 reflects the predictive accuracy of the model from leave-one-out cross validation.

Glucose release can be predicted by as few as three wood chemical composition traits. Next, to identify the smallest possible set of wood traits necessary to predict TWG and glucose release with significant accuracy (Q2 ≥ 0.5), we compared linear models and the more complex machine learning algorithms Generalized Additive Models (GAMs; Hastie and Tibshirani, 1986, Wood, 2006) and Random Forests (Breiman, 2001). For either glucose release or TWG, we generated at least 30 models of each type, and selected the model from each type showing the best prediction accuracy (Dataset S2).

Linear modelling appeared sufficient to predict glucose release with significant accuracy (Fig. 1c; Dataset S2). Interestingly, predictions of glucose release (Q2 = 0.89) relied only on three traits

(Table 1; Dataset S2): G-lignin content, rhamnose content, and 4-O-methyl glucuronic acid content, all of which are related to the chemical composition of the wood, rather than to wood anatomy and structure.

Table 1: Parameters of the linear model predicting glucose released after pretreatment

Trait predicting glucose released in enzymatic hydrolysate after pretreatment	Coefficient/weight of the trait in the linear model predicting glucose release
G-lignin content (G)	-0.007
Rhamnose content (Rha)	-0.035
4-O-methyl glucuronic acid (4-O-MeGlcA)	-0.011
Intercept	0.533

A set of 22 chemical and structural traits predicts TWG. TWG is a composite trait combining glucose release, stem height, stem diameter and wood density, meaning that predicting individually these four components of TWG is required to better predict and interpret TWG (Escamez et al., 2017). We generated models (Dataset S2) that could accurately predict glucose release (Q2 = 0.89, see linear model above), wood density (Q2 = 0.96, GAM), stem height (Q2 = 0.69, GAM), and stem diameter (Q2 = 0.65, Random Forest), leading to a composite model predicting TWG from 22 traits (Q2 = 0.5; Table 2; Dataset S2).

In a previous study on a population of genetically engineered hybrid aspens (BioImprove collection), we had also relied on 22 wood traits to predict TWG (Escamez et al., 2017). This previous model (BioImprove model) could predict TWG of the SwAsp collection, but with lower accuracy (Q2 = 0.32; Fig. S3) than the model developed here with the SwAsp wood properties (Q2 = 0.5; Fig. 1d). The models from both studies shared 11 traits, eight of which had a similar direction of association (positive or negative) to TWG (Table S1). These traits associated with TWG between two very different experimental settings (different species, very different growth conditions and age of the trees) could represent general diagnostic traits for superior biorefinery feedstocks in *Populus* species.

Predictor traits for bioprocessing yield can be genetically uncoupled. We estimated the broad sense heritability (H2) of all the measured wood properties, growth traits and saccharification (Table S2). Some traits, especially linked to xylose content and xylose released by saccharification, showed nearly no heritability, while traits related to tree growth and wood anatomy showed moderate to high heritability (H2 > 0.5). Wood chemical composition traits showed variable heritability, which was generally lower for monosaccharides, such as glucose release predictors rhamnose and 4-O-methyleglucuonic acid content, and higher for lignin composition traits, especially S-type and G-type lignin content (Table S2).

The traits predictive of TWG also displayed a wide range of heritability (Table S2). Most of the predictive traits for TWG were not genetically correlated (Fig. S4), except for obviously related traits (e.g. vessel major diameter and vessel minor diameter). Hence, these traits could be modulated at the genetic level independently of one another, allowing the tailoring of genetic improvement of *Populus* biorefinery feedstocks.

Table 2: Traits required by the composite model predicting total-wood glucose yield (TWG†)

Traits predicting TWG (traits contributing to either of the models for stem height, stem diameter, wood density or glucose release after pretreatment)	**Impact of the trait on TWG in the composite model**
Proportion of carbohydrates (C)	Positive ‡
Proportion of G-lignin (G)	Negative
Proportion of S-lignin (S)	Positive
Proportion of unidentified phenolics from lignin (P)	Negative §
Proportion of total lignin (L)	Positive ¶
Arabinose content (Ara)	Negative
Rhamnose content (Rha)	Negative
Fucose content (Fuc)	Negative §
Mannose content (Man)	Negative §
4-O-methylglucuronic acid content (4-O-meGlcA)	Negative
Extractable/non-crystalline glucose content (Glc)	Positive §
Glucuronic acid content (GlcA)	Negative §
Coarseness	Positive
Cell wall thickness	Positive
Major diameter of vessels (VessMajorDiam)	Positive
Minor diameter of vessels (VessMinorDiam)	Positive §
Vessel cross-sectional perimeter (VessPerim)	Positive
Fraction of the wood area made of fibers (FibFrac)	Positive
Fraction of the wood area made of Vessels (VessFrac)	Negative §
Number of fibers per wood area unit (FibPop)	Positive
Number of vessels per wood area unit (VessPop)	Negative §
Ratio of fibers to vessels (FibPerVess)	Positive §

†TWG (total-wood glucose yield) estimates the glucose released from saccharification after pretreatment for the entire wood biomass of a tree (i.e. taking into account tree growth).

‡This trait's relationship to TWG is non-monotonic (i.e. the direction, positive or negative, is not constant). As a result, the direction (positive or negative) that is observed for a range of values that encompasses a majority of the SwAsp genotypes is indicated.

§Traits that were solely used as predictors for stem diameter could not be given a direction of association from the type of model used (random forest machine learning algorithm). Instead, the direction displayed here for these traits is the relation from pairwise correlation between the predictor trait and the predicted stem diameter.

¶This trait's pairwise correlation to TWG and to the individual traits it predicts is actually negative, but the machine learning algorithms found that it best contributed to predictions by a positive association (in a context where it is used along with other variables rather than individually).

Genetic markers are significantly associated with wood traits that predict bioprocessing yield. To further decipher the genetics underlying wood properties and amenability to bioprocessing, we performed a genome wide association study (GWAS; Dataset S3). We identified only a limited number of significant associations between SNPs and traits (Fig. 2a), consistent with generally low effect sizes of the SNPs (Dataset S3), as well as due to the possible effect of stringent statistical correction for multiple comparisons. Nevertheless, we could identify SNPs significantly associated with wood traits (Fig. 2a), especially traits linked to wood chemical composition.

Although fewer saccharification traits than wood properties were significantly associated with SNPs in the GWAS analyses, we could identify one and three SNPs significantly associated (q-value < 0.05) with glucose released after pretreatment and total hexoses released after pretreatment, respectively (Fig. 2a; Dataset S3). However, the effect size of these SNPs on the saccharification traits

(referred to as "beta" in Dataset S3) was in all cases low, meaning that selection of feedstocks based on these SNPs would only enable limited gain in saccharification yield.

We identified 334 SNPs significantly associated with four wood chemical composition traits that predict TWG (Table 2; Fig. 2a; Dataset S3). Among them, G-lignin and rhamnose content were also two of the three traits predicting glucose release after pretreatment (Table 1; Dataset S2). While most of these SNPs also had low effect sizes on the traits they associated with (Dataset S3), their abundance provides more promising information for the selection of biorefinery feedstocks based in genetic information. Such SNPs may also reveal new candidate genes involved in the regulation of wood chemical composition.

The SNPs significantly associated with saccharification predictor G-lignin fall in and around two genes. The SNPs significantly associated with G-lignin were particularly interesting because 14 of the 16 significant SNPs were located in the body and the regulatory regions of a single gene, Potra000716g05617 (Fig. 2b,c, Fig. S5). The two other SNPs significantly associated with G-lignin were located upstream of the neighbouring gene, Potra000716g05618 (Fig. 2d, Fig. S5). These two genes therefore represent potential targets for the modulation of G- lignin as well as saccharification yield in *Populus* trees.

Both genes encode proteins that have orthologs in other angiosperm plants, including the potential bioenergy feedstocks *Eucalyptus grandis*, *Salix Purpurea* and *Medicago truncatula* (Fig. S6A,B). Potra000716g05617 also has orthologs in the early tracheophyte *Selaginella moellendorffii*, in the moss *Physcomitrella patens* and in the early vascular plant *Amborella trichopoda*, whereas Potra000716g05618 does not (Fig. S6a,b). This suggests that only Potra000716g05617 is conserved among land plants, while Potra000716g05618 would have appeared later, during vascular plant evolution.

The closest homolog of Potra000716g05617 in *Arabidopsis thaliana* is AT3G09210 (Fig. S6a), isolated as part of a protein complex regulating transcription in plastids (plastid transcriptionally active chormose protein [pTAC]13; Pfalz et al., 2006). Potra000716g05618 is annotated as belonging to the family of cytochromes P450 (CYPs) monooxygenases that includes several monolignol biosynthetic genes (Gou et al., 2018), closest to the uncharacterized *Arabidopsis* CYP76G1/AT3G52970 (Höfer et al., 2014; Fig. S6b). Therefore, we hereafter refer to the Potra000716g05618 gene as *Potra-CYP76G1*, and to the Potra000716g05617 gene as *Potra-pTAC13*.

Of the 14 significant SNPs in *Potra-pTAC13*, 13 appeared in perfect linkage disequilibrium (LD), meaning that they all showed the same allele (marked by asterisks in Fig. 2c): either all being homozygous for the major allele, or all being heterozygous, within any one clone. Notably, the minor allele for these SNPs was not found as homozygous in the SwAsp collection. The 14th SNP, situated over 1Kbp upstream of the gene, displayed all three possible allelic variants (homozygous major allele, homozygous minor allele, or heterozygous), while still showing a rather high association with the 13 other SNPs (LD = 0.87). The two remaining significant SNPs, located upstream of *Potra-CYP76G1* (Fig. 2d), showed perfect linkage disequilibrium with the aforementioned 14th SNP (upstream of *Potra-pTAC13*, Fig. 2c), and therefore also a high association (LD = 0.87) with the other 13 significant SNPs.

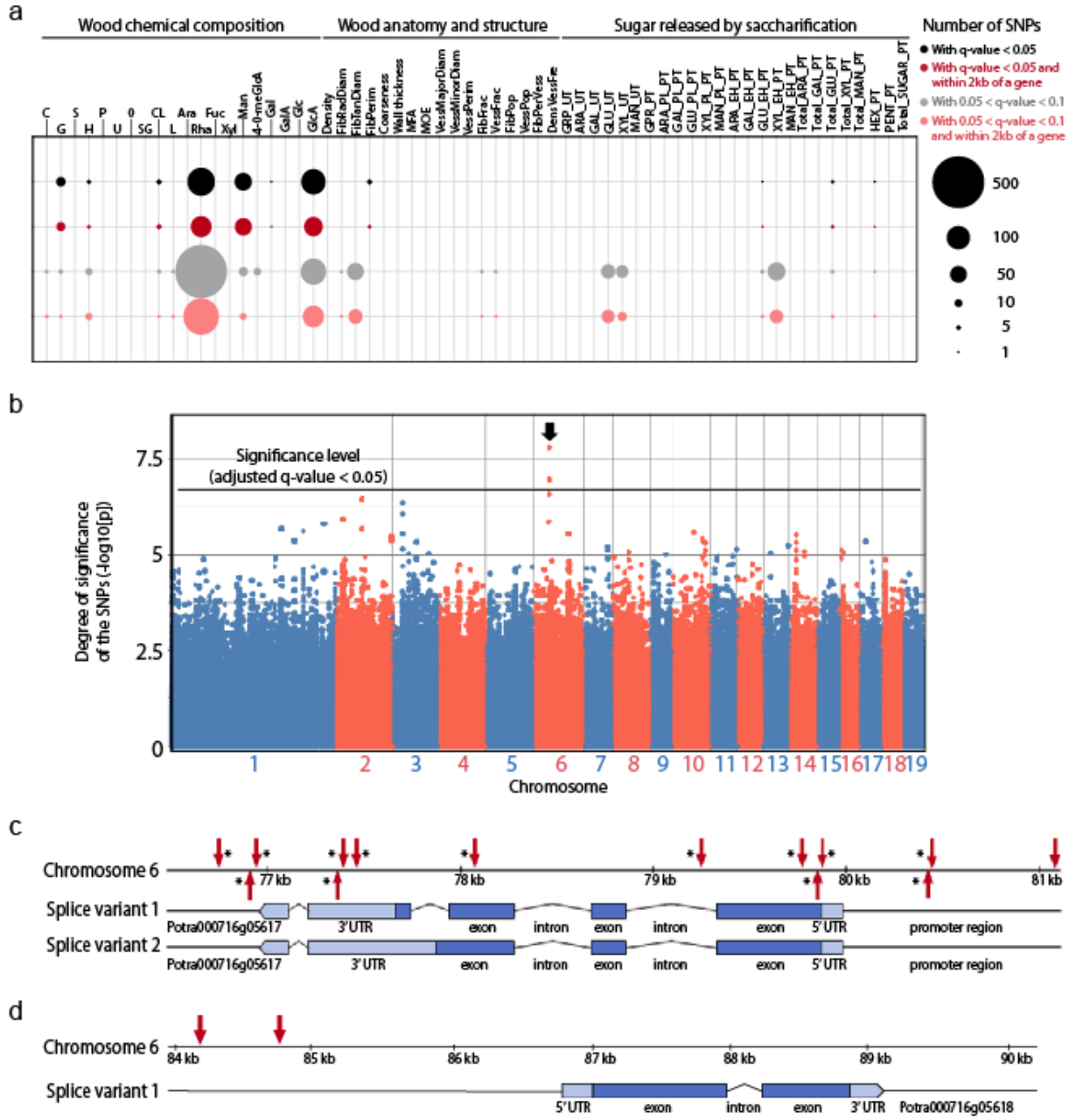

Figure 2. Genome Wide Association Study (GWAS) reveals Single Nucleotide Polymorphisms (SNPs) significantly linked to wood properties.

(a) Number of significant associations (q-value < 0.05) and nearly significant as well as significant associations (q-value < 0.10) between SNPs and the indicated wood traits. SNPs laying in and around (within 2 kb) gene bodies are distinguished from clearly intergenic SNPs (at least 2 kb from any gene). Abbreviations for trait names are clarified in Dataset S1.

(b) Manhattan plot showing the genomic location of the SNPs associated with G-lignin content (X-axis) as well as the degree of significance of the association with G-lignin content (Y-axis). Nearly all the SNPs significantly associated with G-lignin content laid in the same region (arrow), in and around two genes, described in (c) and (d).

(c) Schematic representation of the genomic location of 14 out of the 16 SNPs significantly associated with G-lignin content (red arrows), all situated in and around the uncharacterized Potra000716g05617 gene. Asterisks indicate SNPs in perfect linkage disequilibrium with each-other (i.e. SNPs that always show the same combination of major vs minor allele).

(d) Schematic representation of the genomic location of the other two (out of 16) SNPs significantly associated with G-lignin content (red arrows), both situated upstream of the uncharacterized Potra000716g05618 gene.

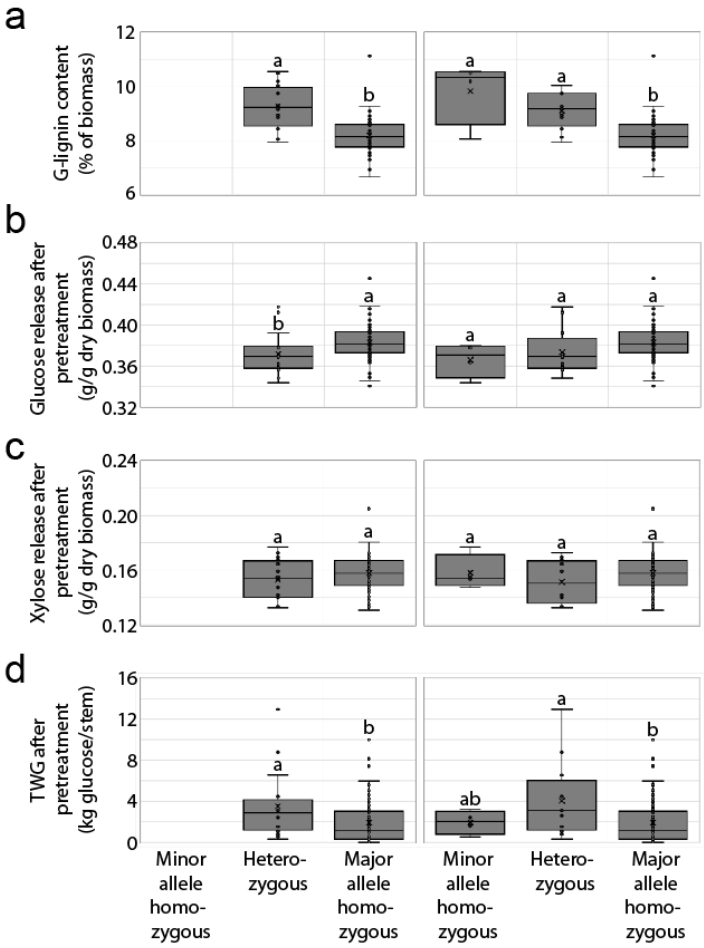

Figure 3. Allelic variation linked to *Potra-pTAC13* and *Potra-CYP76G1* correlates with G-lignin content.
a) Proportion of G-lignin in *Populus* trees in relation to their genotype for the significant SNPs located in the *Potra-pTAC13/Potra-CYP76G1* locus. The left chart represents the 13 SNPs in perfect linkage disequilibrium, while the right chart represents the 14th SNP associated with *Potra-pTAC13* and the two SNPs upstream of *Potra-CYP76G1*. For the location of the SNPs, see Fig. 2 and Fig. S5.
(b) Proportion of glucose released after pretreatment from the wood of *Populus* trees in relation to their genotype for the significant SNPs located in the *Potra-pTAC13/Potra-CYP76G1* locus. The left chart represents the 13 SNPs in perfect disequilibrium, while the right chart represents the 14th SNP associated with *Potra-pTAC13*, as well as the two SNPs upstream of *Potra-CYP76G1*.
(c) Proportion of xylose released after pretreatment from the wood of *Populus* trees in relation to their genotype for the significant SNPs located in the *Potra-pTAC13/ Potra-CYP76G1* locus. The left chart represents the 13 SNPs in perfect disequilibrium, while the right chart represents the 14th SNP associated with *Potra-pTAC13*, as well as the two SNPs upstream of *Potra-CYP76G1*.
(d) Total-wood glucose yield (TWG) after pretreatment from the wood of *Populus* trees in relation to their genotype for the significant SNPs located in the *Potra-pTAC13/Potra-CYP76G1* locus. The left chart represents the 13 SNPs in perfect disequilibrium, while the right chart represents the 14th SNP associated with *Potra-pTAC13*, as well as the two SNPs upstream of *Potra-CYP76G1*.
In all charts, black dots represent individual genotypes. Boxes in the box plots that do not share any letter indicate statistically significant differences. Genotypes that do not share any letter are significantly different ($p < 0.05$)

Allelic variation in all significant SNPs correlates with G-lignin, glucose release and TWG. As expected from the GWAS results linking 16 SNPs and G-lignin, the wood G-lignin content was significantly different for trees displaying different alleles (Fig. 3a): For the 13 *Potra-pTAC13* SNPs in perfect LD, trees harbouring the homozygous major allele contained ~12% less G-lignin than heterozygous trees. Similarly, for the 14th *Potra-pTAC13* SNP and the two *Potra-CYP76G1* SNPs, G-lignin content correlated negatively with the presence of the major allele.

Consistent with the fact that G-lignin is a negative predictor for glucose release after enzymatic hydrolysis (Table 1), allelic variation in the 16 significant SNPs correlated with glucose release in the opposite way as for G-lignin (Fig. 3b). For instance, homozygous trees for the major allele of the 13 SNPs in perfect LD, which correlated with lower G-lignin, showed higher glucose release (Fig. 3b). For the 14th *Potra-pTAC13* SNP as well as the two SNPs upstream of *Potra-CYP76G1*, the glucose release tended to increase with the presence of the major allele variant, although the differences were not significant (Fig. 3b). The relationship between the G-lignin-associated SNPs and glucose release seemed specific, as the allelic variation did not correlate with the release of xylose (Fig. 3c), the second most abundant sugar released from wood by saccharification.

In addition to correlating with G-lignin and glucose release, allelic variation for the 13 *Potra-pTAC13* SNPs in perfect LD correlated significantly with total-wood glucose yield (TWG) such that presence of the minor allele was associated with a significant increase in TWG (Fig. 3d). The association between TWG and the 14th *Potra-pTAC13* SNP or the two SNPs upstream of *Potra-CYP76G1* was less clear, as heterozygous trees displayed higher TWG than homozygous trees for either the major or minor allele (Fig. 3d). These observations therefore strongly suggest that only the 13 SNPs in perfect

LD represent genetic markers for higher yield, in line with the fact that they predict G-lignin, which is itself a predictor trait for yield.

Variation in *Potra-pTAC13* expression correlates with both allelic variation and G-lignin content. To understand how *Potra-pTAC13* may relate to wood G-lignin content and bioprocessing yield, we investigated the possibility that its variants might result in differential levels of expression. An alternative hypothesis would be a change in protein sequence, but all the SNPs in the *Potra-pTAC13* coding sequence either resulted in silent mutations, except for amino acid 316 (of 360) which undergoes a conservative substitution (substitution for an amino-acid with similar biochemical properties, here glycine for the minor allele and serine for the major allele). Hence, a change in gene expression seems more promising a hypothesis than a change in protein sequence.

To test this hypothesis, we selected a subset of trees from the SwAsp collection that covers the range of allelic variation as well as a range of G-lignin content. The corresponding trees were clonally propagated *in vitro*, and the basal part of three replicates per genotype were harvested for measuring *Potra-pATC13* transcript levels by qPCR, while also analysing G-lignin content from the same samples by pyrolysis-GC/MS. The expression of *Potra-pTAC13* apparently correlated with the presence of the major allele of the 13 SNPs in perfect linkage disequilibrium (Fig. 4a), although not significantly (p-value = 0.13). For the 14th SNP, further upstream in the promoter region of *Potra-pTAC13*, homozygous genotypes for the major allele showed significantly higher gene expression compared with the genotypes bearing one or two copies of the minor allele (Fig. 4a). The G-lignin content of these in-vitro grown trees showed a trend towards a negative correlation (r = -0.28; p-value = 0.15) with the *Potra-pTAC13* transcript levels (Fig. 4b). It is therefore possible that the accumulation of G-lignin is negatively influenced by the level of *Potra-pTAC13* expression.

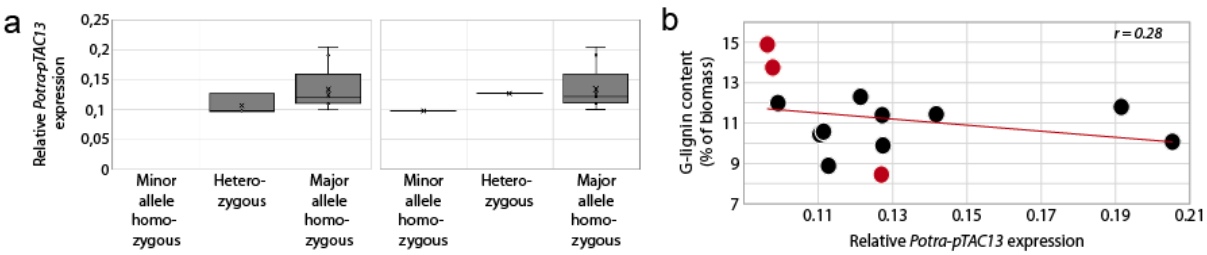

Figure 4. Allelic variants of *Potra-pTAC13* display differential expression correlated with G-lignin content.
(a) Relative expression of *Potra-pTAC13* in a subset of the SwAsp *Populus* trees in relation to their genotype for the significant SNPs located in the *Potra-pTAC13* gene. The left chart represents the 13 SNPs in perfect linkage disequilibrium, while the right chart represents the 14th SNP.
Black dots represent individual genotypes. Boxes in the box plots that do not share any letter indicate statistically significant differences.
(b) G-lignin content in relation to transcript levels of *Potra-pTAC13*. Red-colored dots represent heterozygous trees for the 13 SNPs in perfect linkage disequilibrium, while black dots represent homozygous trees for the major allele.

Discussion

Wood biomass from fast growing trees represents a promising source of biofuels and other bioproducts to transition away from petroleum (Percival Zhang, 2013, Ragauskas et al., 2014). The high cost of deconstructing woody biomass, however, hinders wood refining into biofuels and other bioproducts (McCann and Carpita, 2015). To overcome this biomass recalcitrance, it is necessary to understand how wood properties relate to wood recalcitrance (Escamez et al., 2017). Similar to our previous approach in a population of genetically engineered hybrid aspens (Escamez et al., 2017), we now report the phenotyping of a population of European aspen natural genotypes for 65 traits related

to tree growth, wood anatomy and structure, wood cell wall chemical composition, and wood bioprocessing yield.

Using statistical modelling and machine learning algorithms, we identified a set of wood traits that could predict glucose yield from saccharification, as well as the estimated total glucose yield from the entire stem wood of the trees (TWG). Eight of the predictor traits for yield showed similar association with TWG as found in a previous study in a collection of transgenic hybrid aspens (Escamez et al., 2017; Table S2), despite monitoring different trees grown in very different conditions (two months in greenhouses in the previous study versus ten years in a plantation in this study). These traits may therefore represent diagnostic traits for important bioprocessing properties. Interestingly, several of these diagnostic traits showed rather high heritability (Table S2), while also displaying very limited genetic correlation with each other (Fig. S2), suggesting that genetic gains could be tailored towards individual traits based on desired effects.

Identifying the genetics underlying wood properties that foster bioprocessing potential should greatly help with selecting or creating superior biorefinery feedstocks (Fahrenkrog et al., 2017). While in theory GWAS is a promising strategy to identify single nucleotide polymorphisms associated with traits of interest (Tuskan et al., 2019), GWAS approaches have often found only few genetic variants associated with wood properties. Previous attempts to optimize GWAS in tree species have consisted in data-mining approaches such as reducing the genomic space under investigation to genomic areas previously associated with wood (Porth et al., 2013), or combining traits into multi-trait phenotypes that may reveal associations with pleiotropic genes (Chhetri et al., 2019, Chhetri et al., 2020). In this study, we employ another approach: decomposing complex traits into better defined traits that are more likely to relate to specific loci within the genome.

We indeed found that finer traits, such as wood cell wall content in a specific monosugar or a specific sub-type of lignin, were associated with more SNPs than composite traits, such as total lignin content or sugar release after pretreatment (Fig. 2a). Importantly, while we could only find few statistically significant associations with important bioprocessing yield, we could find many more significant associations with cell wall chemical composition traits that predict bioprocessing yield (e.g. 146 SNPs with q-value < 0.05 for rhamnose content, or 16 SNPs for G-lignin). G-lignin content represents an interesting example because looking specifically at the SNPs significantly associated with G-lignin showed that most of them also correlated with significant differences in glucose release after pretreatment and with TWG (Fig. 3b,d), something that the GWAS itself had not revealed. Hence, SNPs for composite traits can be indirectly identified through breaking down these broadly defined traits into finer traits that predict them. This approach therefore represents a complement to the direct search for associations by GWAS.

While we found significant associations with new candidate genes for numerous wood cell wall compositional traits, we found very few significant associations with known biosynthetic genes for these cell wall components. This may seem surprising, but such lack of significant association with known biosynthetic genes has been common in previous GWAS work for wood chemical composition traits (Guerra et al., 2019).

Statistical significance of an association usually scales with the effect size of the SNP. If a trait relies on the function of many genes, SNPs associated with that trait tend to have small effect sizes, and there tends to be fewer significant associations overall. Many genes are known to partake in the biosynthetic functions for the main wood cell wall chemical components lignin, hemicelluloses, and cellulose (Goujon et al., 2003, Somerville, 2006, Scheller and Ulvskov, 2010). This reliance on large sets of genes could in part explain why so few significant SNPs were found within or near known biosynthetic genes for these important lignocellulose components.

Alternatively, or in addition, it is possible that some of these biosynthetic pathways currently undergo selective pressure that would lead to a lower proportion of SNPs in the corresponding genes. For instance, while the average number of SNPs in and around (2kbp) of genes in the SwAsp population is 104.5, the average number of SNPs in and around known or suspected lignin biosynthetic genes is only 82.2, while for hemicelluloses biosynthetic genes it is 104.5, similar to the genome-wide average, and 150.6 in cellulose biosynthetic genes (Dataset S4). Therefore, the SwAsp population may have undergone recent selective pressure on the lignin biosynthetic pathway, resulting in fewer SNPs in known biosynthetic genes.

Even when performing association on a gene space limited to wood-expressed genes, Porth et al (2013) found mostly associations with SNPs in genes that had not previously been linked to wood formation. This raises the question of whether the already characterized biosynthetic genes for wood cell wall components would be so essential that they undergo strong selective pressure in general, and not just in the SwAsp population, forbidding observation of extensive natural variation. If this were the case, then GWAS approaches for wood cell wall composition would mostly reveal associations with genes that, while less essential than core regulators of cell wall biosynthesis, are still significant contributors to the quantitative changes that can be naturally sustained. These sort of genes, capable of quantitatively modulating a trait of interest, while not being essential to biological functions of this trait, might be seen as ideal targets for feedstocks improvements. Indeed, targeting these genes for selection or genetic engineering would guarantee that the feedstocks would remain adapted to a wide range of environments, while displaying quantitative improvements in traits facilitating downstream utilization of the biomass.

Interestingly, all of the 16 SNPs significantly associated with G-lignin content clustered at the same genomic region (Fig. 2b). All 16 SNPs laid in and around two genes (Fig. 2c,d) that had not been previously associated with cell walls or lignin. One of these genes, *Potra-CYP76G1*, encodes for a cytochrome P450 hydroxylase which is implicated in flavonoid biosynthesis on the basis of sequence homology, and could therefore be involved in monolignol biosynthesis, or indirectly regulate lignin biosynthesis, for example by competing with monolignol biosynthesis for substrates such as hydroxyl groups. The other gene, *Potra-pTAC13*, is thought to be involved in regulation of transcription in plastids (Pfalz et al., 2006), although other functions cannot be excluded. Even if *Potra-pTAC13* functioned solely in plastids, it could still indirectly regulate lignin biosynthesis by modulating the metabolic flux through the plastidial shikimate pathway (as in Eudes et al., 2015), which ultimately yields phenylalanine, the amino-acid precursor of monolignols. Future studies will hopefully reveal the role of these new candidate regulators of lignin chemical composition, especially as allelic variation associated with at least *Potra-pTAC13* also correlates with bioprocessing yield (Fig. 3b,d).

The allelic variants we identified as significantly (q-value < 0.05) associated with G-lignin content, in and around *Potra-pTAC13* as well as upstream of *Potra-CYP76G1*, all showed significant correlations with TWG (Fig. 4d), and they could therefore be integrated into a set of genetic markers for superior biorefinery aspen feedstocks. This is consistent with the fact that G-lignin is a predictor of both TWG as well as glucose release from saccharification. The other traits that predict glucose release from saccharification are rhamnose content and 4-O-methyl glucuronic acid content. We identified significant associations between these traits and up to 668 and 10 SNPs, respectively (considering q-value < 0.1; Fig. 2a; Dataset S3), or at least 146 SNPs for rhamnose content (considering q-value < 0.05), which could also be added in a set of genetic markers for the selection of superior feedstocks. Considering that we had only identified three (q-value < 0.05) to seven (q-value < 0.1) SNPs significantly associated with glucose release directly (Fig. 2a; Dataset S3), our approach of "indirect" associations

via predictor traits allowed increasing the number of potential genetic markers for glucose release by an order of magnitude.

Author Contribution

HT designed the study, with help from KMR, SJ, GS, LGS, LJJ, NS, and SE. Phenotypic characterization of the trees was performed by SE, ML, MLG, KMR, and TG, with supervision by LJJ, GS, and HT. Statistical modelling of relationships between traits was performed by SE. LGS performed analyses of broad-sense heritability and genetic correlations. Genome-wide association study for identification of SNPs was performed by NM and NS. Gene expression analyses were performed by SE and ML. SE and HT wrote the manuscript, with assistance from all co-authors.

Acknowledgements

The authors thank the UPSC Biopolymer Analytical Platform (supported by Bio4Energy and TC4F) and its manager, Junko Takahashi-Schmidt, for the analyses of the wood chemical composition traits. We thank Veronica Bourquin and Marlene Karlsson for help in preparing the wood samples for analyses.
This work was supported by grants from Formas (942-2015-84 and 2018-01381), the Knut and Alice Wallenberg Foundation (2016.0341 and 2016.0352), and the Swedish Governmental Agency for Innovation Systems VINNOVA (2016-00504). LJJ and MLG also acknowledge financial support from the strategic research initiative Bio4Energy (www.bio4energy.se).

Data Availability

All the data used for analyses in this manuscript is either displayed in the supplementary datasets, or available upon request to the corresponding author.

References

BAR-ON, Y. M., PHILLIPS, R. & MILO, R. 2018. The biomass distribution on Earth. *Proceedings of the National Academy of Sciences,* 115**,** 6506-6511.

BOERJAN, W., RALPH, J. & BAUCHER, M. 2003. Lignin biosynthesis. *Annual Review of Plant Biology,* 54**,** 519-546.

BREIMAN, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

CARMER, S. G. & SWANSON, M. R. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association,* 68**,** 66-74.

CHHETRI, H. B., FURCHES, A., MACAYA-SANZ, D., WALKER, A. R., KAINER, D., JONES, P., HARMAN-WARE, A. E., TSCHAPLINSKI, T. J., JACOBSON, D. & TUSKAN, G. A. 2020. Genome-Wide Association Study of Wood Anatomical and Morphological Traits in Populus trichocarpa. *Frontiers in plant science,* 11**,** 1391.

CHHETRI, H. B., MACAYA-SANZ, D., KAINER, D., BISWAL, A. K., EVANS, L. M., CHEN, J. G., COLLINS, C., HUNT, K., MOHANTY, S. S., ROSENSTIEL, T., et al. 2019. Multitrait genome-wide association analysis of Populus trichocarpa identifies key polymorphisms controlling morphological and physiological traits. *New Phytologist,* 223**,** 293-309.

DICKMANN, D. I. 2006. Silviculture and biology of short-rotation woody crops in temperate regions: Then and now. *Biomass and Bioenergy,* 30**,** 696-705.

DU, Q., LU, W., QUAN, M., XIAO, L., SONG, F., LI, P., ZHOU, D., XIE, J., WANG, L. & ZHANG, D. 2018. Genome-wide association studies to improve wood properties: challenges and prospects. *Frontiers in plant science,* 9.

ESCAMEZ, S., LATHA GANDLA, M., DERBA-MACELUCH, M., LUNDQVIST, S.-O., MELLEROWICZ, E. J., JÖNSSON, L. J. & TUOMINEN, H. 2017. A collection of genetically engineered Populus trees reveals wood biomass traits that predict glucose yield from enzymatic hydrolysis. *Scientific Reports,* 7**,** 15798.

EUDES, A., SATHITSUKSANOH, N., BAIDOO, E. E., GEORGE, A., LIANG, Y., YANG, F., SINGH, S., KEASLING, J. D., SIMMONS, B. A. & LOQUÉ, D. 2015. Expression of a bacterial 3-dehydroshikimate dehydratase reduces lignin content and improves biomass saccharification efficiency. *Plant biotechnology journal,* 13**,** 1241-1250.

FAHRENKROG, A. M., NEVES, L. G., RESENDE, M. F., VAZQUEZ, A. I., CAMPOS, G., DERVINIS, C., SYKES, R., DAVIS, M., DAVENPORT, R., BARBAZUK, W. B., et al. 2017. Genome-wide association study reveals putative regulators of bioenergy traits in Populus deltoides. *New Phytologist,* 213**,** 799-811.

FURBANK, R. T. & TESTER, M. 2011. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends in plant science,* 16**,** 635-644.

GANDLA, M. L., DERBA-MACELUCH, M., LIU, X., GERBER, L., MASTER, E. R., MELLEROWICZ, E. J. & JÖNSSON, L. J. 2015. Expression of a fungal glucuronoyl esterase in Populus: Effects on wood properties and saccharification efficiency. *Phytochemistry,* 112**,** 210-220.

GERBER, L., ELIASSON, M., TRYGG, J., MORITZ, T. & SUNDBERG, B. 2012. Multivariate curve resolution provides a high-throughput data processing pipeline for pyrolysis-gas chromatography/mass spectrometry. *Journal of Analytical and Applied Pyrolysis,* 95**,** 95-100.

GILMOUR, A., THOMPSON, R., CULLIS, B. & WELHAM, S. 1997. ASREML [computer program]. *NSW Agriculture, Orange, Australia*.

GOU, M., RAN, X., MARTIN, D. W. & LIU, C.-J. 2018. The scaffold proteins of lignin biosynthetic cytochrome P450 enzymes. *Nature plants,* 4**,** 299-310.

GOUJON, T., SIBOUT, R., EUDES, A., MACKAY, J. & JOUANIN, L. 2003. Genes involved in the biosynthesis of lignin precursors in Arabidopsis thaliana. *Plant Physiology and Biochemistry,* 41**,** 677-687.

GRIMBERG, Å., LAGER, I., STREET, N. R., ROBINSON, K. M., MARTTILA, S., MÄHLER, N., INGVARSSON, P. K. & BHALERAO, R. P. 2018. Storage lipid accumulation is controlled by photoperiodic signal

acting via regulators of growth cessation and dormancy in hybrid aspen. *New Phytologist,* 219**,** 619-630.

GUERRA, F. P., SUREN, H., HOLLIDAY, J., RICHARDS, J. H., FIEHN, O., FAMULA, R., STANTON, B. J., SHUREN, R., SYKES, R., DAVIS, M. F., et al. 2019. Exome resequencing and GWAS for growth, ecophysiology, and chemical and metabolomic composition of wood of Populus trichocarpa. *BMC genomics,* 20**,** 1-14.

GUERRA, F. P., WEGRZYN, J. L., SYKES, R., DAVIS, M. F., STANTON, B. J. & NEALE, D. B. 2013. Association genetics of chemical wood properties in black poplar (Populus nigra). *New Phytologist,* 197**,** 162-176.

HASTIE, T. & TIBSHIRANI, R. 1986. Generalized additive models. *Statistical science,* 1**,** 297-310.

HOU, Z., LI, A. & ZHANG, J. 2020. Genetic architecture, demographic history, and genomic differentiation of Populus davidiana revealed by whole-genome resequencing. *Evolutionary applications,* 13**,** 2582-2596.

HÖFER, R., BOACHON, B., RENAULT, H., GAVIRA, C., MIESCH, L., IGLESIAS, J., GINGLINGER, J.-F., ALLOUCHE, L., MIESCH, M., GREC, S., et al. 2014. Dual function of the cytochrome P450 CYP76 family from Arabidopsis thaliana in the metabolism of monoterpenols and phenylurea herbicides. *Plant physiology,* 166**,** 1149-1161.

LIN, Y.-C., WANG, J., DELHOMME, N., SCHIFFTHALER, B., SUNDSTRÖM, G., ZUCCOLO, A., NYSTEDT, B., HVIDSTEN, T. R., DE LA TORRE, A., COSSU, R. M., et al. 2018. Functional and evolutionary genomic inferences in Populus through genome and population sequencing of American and European aspen. *Proceedings of the National Academy of Sciences,* 115**,** E10970-E10978.

LIU, Y.-J., WANG, X.-R. & ZENG, Q.-Y. 2019. De novo assembly of white poplar genome and genetic diversity of white poplar population in Irtysh River basin in China. *Science China Life Sciences,* 62**,** 609-618.

LIVAK, K. J. & SCHMITTGEN, T. D. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. *methods,* 25**,** 402-408.

LUNDQVIST, S.-O., OLSSON, L., EVANS, R., CHEN, F. F. & VAPAAVUORI, E. 2010. Variations in properties of hardwood analysed with SilviScan -- Examples of wood, fibre and vessel properties of birch (*Betula*). *The 4th Conference on Hardwood Research and Utilisation in Europe 2010.,* Hardwood Science and Technology.

LUQUEZ, V., HALL, D., ALBRECTSEN, B. R., KARLSSON, J., INGVARSSON, P. & JANSSON, S. 2008. Natural phenological variation in aspen (Populus tremula): the SwAsp collection. *Tree Genetics & Genomes,* 4**,** 279-292.

MA, T., WANG, J., ZHOU, G., YUE, Z., HU, Q., CHEN, Y., LIU, B., QIU, Q., WANG, Z., ZHANG, J., et al. 2013. Genomic insights into salt adaptation in a desert poplar. *Nature communications,* 4**,** 1-9.

MCCANN, M. C. & CARPITA, N. C. 2015. Biomass recalcitrance: a multi-scale, multi-factor and conversion-specific property. *Journal of Experimental Botany***,** doi: 10.1093/jxb/erv267.

MENG, X., PU, Y., YOO, C. G., LI, M., BALI, G., PARK, D. Y., GJERSING, E., DAVIS, M. F., MUCHERO, W., TUSKAN, G. A., et al. 2017. An In-Depth Understanding of Biomass Recalcitrance Using Natural Poplar Variants as the Feedstock. *ChemSusChem,* 10**,** 139-150.

MIELENZ, J. R., BARDSLEY, J. S. & WYMAN, C. E. 2009. Fermentation of soybean hulls to ethanol while preserving protein value. *Bioresource Technology,* 100**,** 3532-3539.

MOLA-YUDEGO, B., AREVALO, J., DÍAZ-YÁÑEZ, O., DIMITRIOU, I., HAAPALA, A., FERRAZ FILHO, A. C., SELKIMÄKI, M. & VALBUENA, R. 2017. Wood biomass potentials for energy in northern Europe: Forest or plantations? *Biomass and Bioenergy,* 106**,** 95-103.

MUCHERO, W., GUO, J., DIFAZIO, S. P., CHEN, J.-G., RANJAN, P., SLAVOV, G. T., GUNTER, L. E., JAWDY, S., BRYAN, A. C., SYKES, R. Z., et al. 2015. High-resolution genetic mapping of allelic variants associated with cell wall chemistry in Populus. *BMC Genomics,* 16**,** 1.

MÄHLER, N., SCHIFFTHALER, B., ROBINSON, K. M., TEREBIENIEC, B. K., VUČAK, M., MANNAPPERUMA, C., BAILEY, M. E., JANSSON, S., HVIDSTEN, T. R. & STREET, N. R. 2020. Leaf shape in Populus tremula is a complex, omnigenic trait. *Ecology and evolution,* 10**,** 11922-11940.

MÄHLER, N., WANG, J., TEREBIENIEC, B. K., INGVARSSON, P. K., STREET, N. R. & HVIDSTEN, T. R. 2017. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS genetics,* 13**,** e1006402.

NORDBORG, M. & WEIGEL, D. 2008. Next-generation genetics in plants. *Nature,* 456**,** 720-723.

PERCIVAL ZHANG, Y.-H. 2013. Next generation biorefineries will solve the food, biofuels, and environmental trilemma in the energy–food–water nexus. *Energy Science & Engineering,* 1**,** 27-41.

PFALZ, J., LIERE, K., KANDLBINDER, A., DIETZ, K.-J. & OELMÜLLER, R. 2006. pTAC2,-6, and-12 are components of the transcriptionally active plastid chromosome that are required for plastid gene expression. *The Plant Cell,* 18**,** 176-197.

PORTER, H. F. & O'REILLY, P. F. 2017. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Scientific reports,* 7**,** 1-12.

PORTH, I., KLAPŠTE, J., SKYBA, O., HANNEMANN, J., MCKOWN, A. D., GUY, R. D., DIFAZIO, S. P., MUCHERO, W., RANJAN, P., TUSKAN, G. A., et al. 2013. Genome-wide association mapping for wood characteristics in Populus identifies an array of candidate single nucleotide polymorphisms. *New Phytologist,* 200**,** 710-726.

QIU, D., BAI, S., MA, J., ZHANG, L., SHAO, F., ZHANG, K., YANG, Y., SUN, T., HUANG, J., ZHOU, Y., et al. 2019. The genome of Populus alba x Populus tremula var. glandulosa clone 84K. *DNA Research,* 26**,** 423-431.

RAGAUSKAS, A. J., BECKHAM, G. T., BIDDY, M. J., CHANDRA, R., CHEN, F., DAVIS, M. F., DAVISON, B. H., DIXON, R. A., GILNA, P., KELLER, M., et al. 2014. Lignin valorization: improving lignin processing in the biorefinery. *Science,* 344**,** 1246843.

SANNIGRAHI, P., RAGAUSKAS, A. J. & TUSKAN, G. A. 2010. Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels, Bioproducts and Biorefining,* 4**,** 209-226.

SCHELLER, H. V. & ULVSKOV, P. 2010. Hemicelluloses. *Annual Review of Plant Biology,* 61**,** 263-289.

SJÖDIN, A., STREET, N. R., SANDBERG, G., GUSTAFSSON, P. & JANSSON, S. 2009. The Populus Genome Integrative Explorer (PopGenIE): a new resource for exploring the Populus genome. *New phytologist,* 182**,** 1013-1025.

SOMERVILLE, C. 2006. Cellulose synthesis in higher plants. *Annual Review of Cell and Developmental Biology,* 22**,** 53-78.

STOREY, J. D., BASS, A. J., DABNEY, A. & ROBINSON, D. 2021. qvalue: Q-value estimation for false discovery rate control. *R package version 2.24.0*.

STOREY, J. D. & TIBSHIRANI, R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences,* 100**,** 9440-9445.

SWEELEY, C. C., ELLIOTT, W., FRIES, I. & RYHAGE, R. 1966. Mass spectrometric determination of unresolved components in gas chromatographic effluents. *Analytical chemistry,* 38**,** 1549-1553.

TRYGG, J. & WOLD, S. 2002. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics,* 16**,** 119-128.

TUSKAN, G. A., DIFAZIO, S., JANSSON, S., BOHLMANN, J., GRIGORIEV, I., HELLSTEN, U., PUTNAM, N., RALPH, S., ROMBAUTS, S., SALAMOV, A., et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *science,* 313**,** 1596-1604.

TUSKAN, G. A., MUCHERO, W., TSCHAPLINSKI, T. J. & RAGAUSKAS, A. J. 2019. Population-level approaches reveal novel aspects of lignin biosynthesis, content, composition and structure. *Current opinion in biotechnology,* 56**,** 250-257.

VAN ACKER, R., LEPLÉ, J.-C., AERTS, D., STORME, V., GOEMINNE, G., IVENS, B., LÉGÉE, F., LAPIERRE, C., PIENS, K., VAN MONTAGU, M. C., et al. 2014. Improved saccharification and ethanol yield from field-grown transgenic poplar deficient in cinnamoyl-CoA reductase. *Proceedings of the National Academy of Sciences,* 111**,** 845-850.

VANDESOMPELE, J., DE PRETER, K., PATTYN, F., POPPE, B., VAN ROY, N., DE PAEPE, A. & SPELEMAN, F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology,* 3**,** research0034.

WANG, J., DING, J., TAN, B., ROBINSON, K. M., MICHELSON, I. H., JOHANSSON, A., NYSTEDT, B., SCOFIELD, D. G., NILSSON, O., JANSSON, S., et al. 2018a. A major locus controls local adaptation and adaptive life history variation in a perennial plant. *Genome biology,* 19**,** 72.

WANG, Z., PAWAR, P. M.-A., DERBA-MACELUCH, M., HEDENSTRÖM, M., CHONG, S.-L., TENKANEN, M., JÖNSSON, L. J. & MELLEROWICZ, E. J. 2020. Hybrid aspen expressing a carbohydrate esterase family 5 acetyl xylan esterase under control of a wood-specific promoter shows improved saccharification. *Frontiers in plant science,* 11**,** 380.

WANG, Z., WINESTRAND, S., GILLGREN, T. & JÖNSSON, L. J. 2018b. Chemical and structural factors influencing enzymatic saccharification of wood from aspen, birch and spruce. *Biomass and Bioenergy,* 109**,** 125-134.

WILKERSON, C., MANSFIELD, S., LU, F., WITHERS, S., PARK, J.-Y., KARLEN, S., GONZALES-VIGIL, E., PADMAKSHAN, D., UNDA, F., RENCORET, J. & RALPH, J. 2014. Monolignol ferulate transferase introduces chemically labile linkages into the lignin backbone. *Science,* 344**,** 90-93.

WOOD, S. 2006. *Generalized additive models: an introduction with R*, CRC press.

WULLSCHLEGER, S. D., WESTON, D. J., DIFAZIO, S. P. & TUSKAN, G. A. 2013. Revisiting the sequencing of the first tree genome: Populus trichocarpa. *Tree physiology,* 33**,** 357-364.

XIE, M., MUCHERO, W., BRYAN, A. C., YEE, K. L., GUO, H.-B., ZHANG, J., TSCHAPLINSKI, T., SINGAN, V. R., LINDQUIST, E., PAYYAVULA, R. S., et al. 2018. A 5-enolpyruvylshikimate 3-phosphate synthase functions as a transcriptional repressor in Populus. *The Plant Cell***,** tpc. 00168.2018.

YANG, W., WANG, K., ZHANG, J., MA, J., LIU, J. & MA, T. 2017. The draft genome sequence of a desert tree Populus pruinosa. *Gigascience,* 6**,** gix075.

YOO, C. G., YANG, Y., PU, Y., MENG, X., MUCHERO, W., YEE, K. L., THOMPSON, O. A., RODRIGUEZ, M., BALI, G., ENGLE, N. L., et al. 2017. Insights of biomass recalcitrance in natural Populus trichocarpa variants for biomass conversion. *Green Chemistry,* 19**,** 5467-5478.

ZHANG, Z., CHEN, Y., ZHANG, J., MA, X., LI, Y., LI, M., WANG, D., KANG, M., WU, H., YANG, Y., et al. 2020. Improved genome assembly provides new insights into genome evolution in a desert poplar (Populus euphratica). *Molecular ecology resources,* 20**,** 781-794.

ZHOU, X. & STEPHENS, M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics,* 44**,** 821-824.

ZHOU, X. & STEPHENS, M. 2014. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods,* 11**,** 407-409.

Supplementary data

Fig. S1: Schematic representation of SwAsp tree harvest and sampling.

Fig. S2: Multivariate analysis of the potential relationships between wood properties and glucose release or TWG.

Fig. S3: Comparison of the predictions for TWG using the SwAsp model (from this study) and the formerly developed BioImprove model (Escamez et al., 2017).

Fig. S4: Heatmap of the genetic correlations between traits.

Fig. S5: Genomic location of the two genes associated with G-lignin content.

Fig. S6: Phylogenetic trees of the Potra-pTAC13 and Potra-CYP76G1 homologs in plants.

Dataset S1: Traits measurements on the SwAsp trees for each genotype.

Dataset S2: Best models of each type (multiple linear regression, GAM, and Random Forest) for 8 biorefinery-yield related trait (saccharification or TWG-related).

Dataset S3: Lists of the top 1000 most significant SNPs for each wood trait (chemistry, anatomy, and structure) as well as for each saccharification trait.

Dataset S4: SNP frequency for known cellulose, hemicelluloses, and lignin biosynthetic genes.

Fig. S1

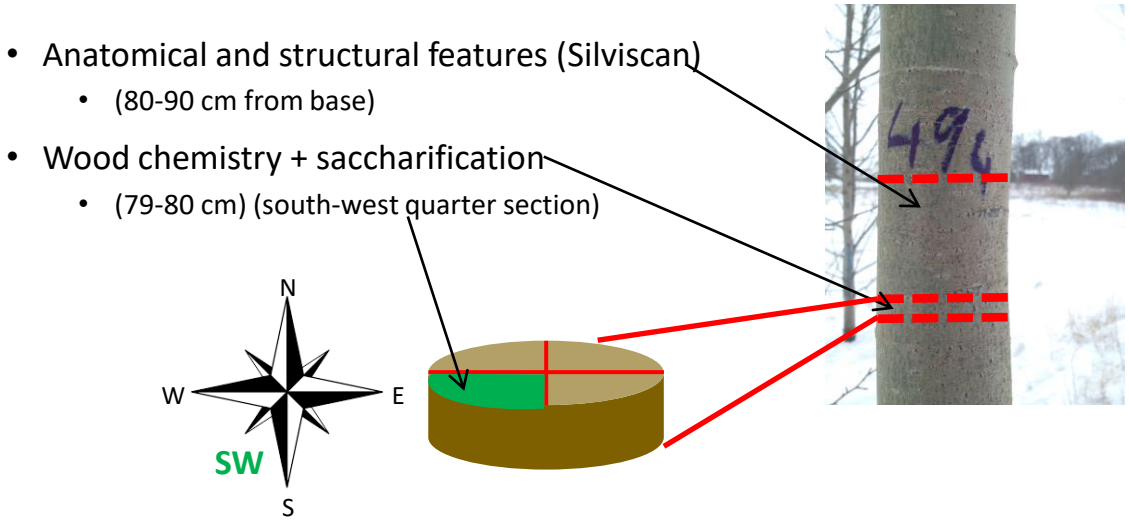

Figure S1.
Schematic representation of SwAsp sample harvest.

Fig. S2

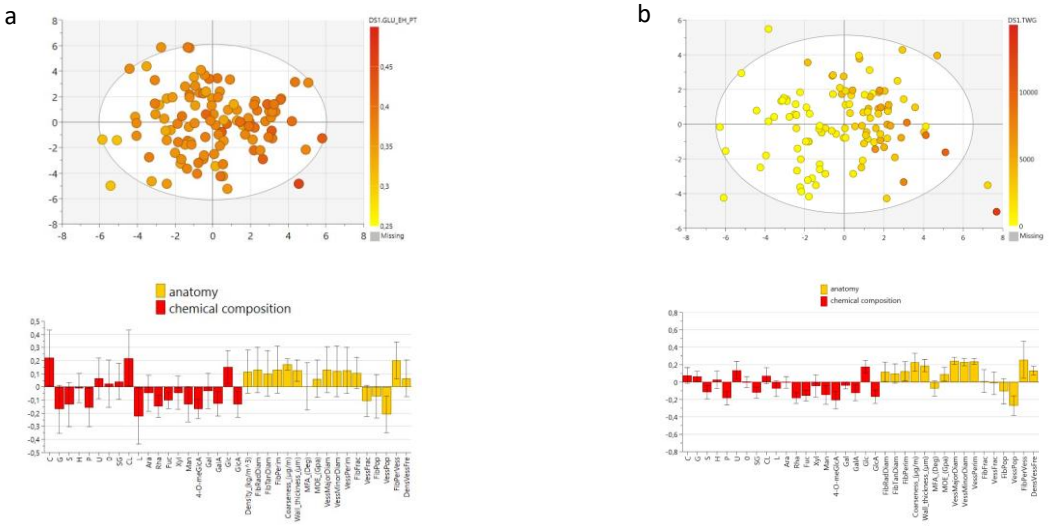

Figure S2.
(a) Orthogonal Projection to Latent Structure (OPLS) analysis relating variation in glucose release after pretreatment between genotypes (up) to wood chemical composition and wood anatomy traits (down). The predictive component separates the lines along the X-axis of the scatter plot, while separation along the Y-axis is not predictive. The dots on the scatter plot (up) correspond to SwAsp genotypes, while their color indicates the median glucose release after pretreatment for each genotype. The bars in the bar chart (bottom) indicate the coefficient ("weight") of each trait in the model.
(b) Orthogonal Projection to Latent Structure (OPLS) analysis relating variation in total-wood glucose yield (TWG) between genotypes (up) to wood chemical composition and wood anatomy traits (down). The predictive component separates the lines along the X-axis of the scatter plot, while separation along the Y-axis is not predictive. The dots on the scatter plot (up) correspond to SwAsp genotypes, while their color indicates the median TWG for each genotype. The bars in the bar chart (bottom) indicate the coefficient ("weight") of each trait in the model.

Fig. S3

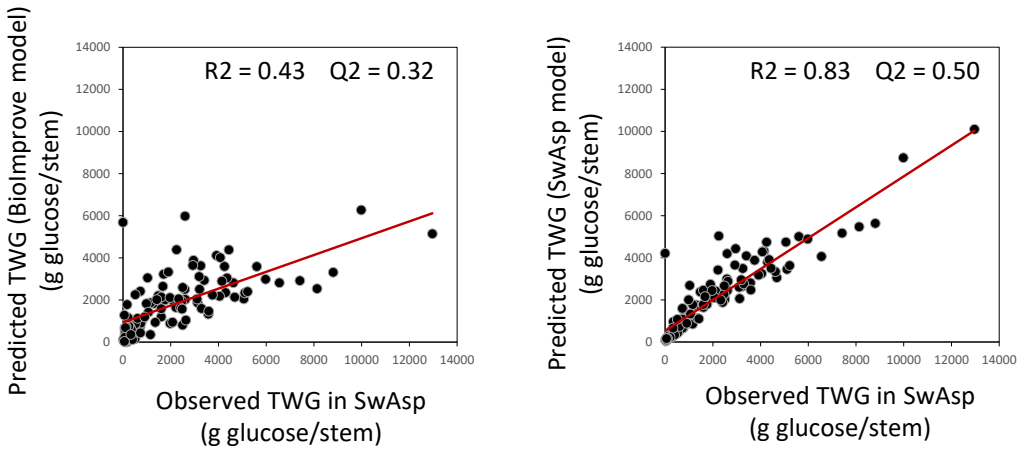

Figure S3.

(a) Scatter plot showing the relatively good correlation between the observed TWG for the SwAsp trees (x-axis), and the predicted TWG (y-axis) based on wood properties using a previously developed statistical model (BioImprove model; Escamez et al., 2017). R2 indicates the variance explained, while Q2 reflects the predictive accuracy of the model from leave-one-out cross validation.

(b) Scatter plot showing the very good correlation between the observed TWG for the SwAsp trees (x-axis), and the predicted TWG (y-axis) based on wood properties using a newly developed statistical model (SwAsp model; this study). R2 indicates the variance explained, while Q2 reflects the predictive accuracy of the model from leave-one-out cross validation.

Fig. S4

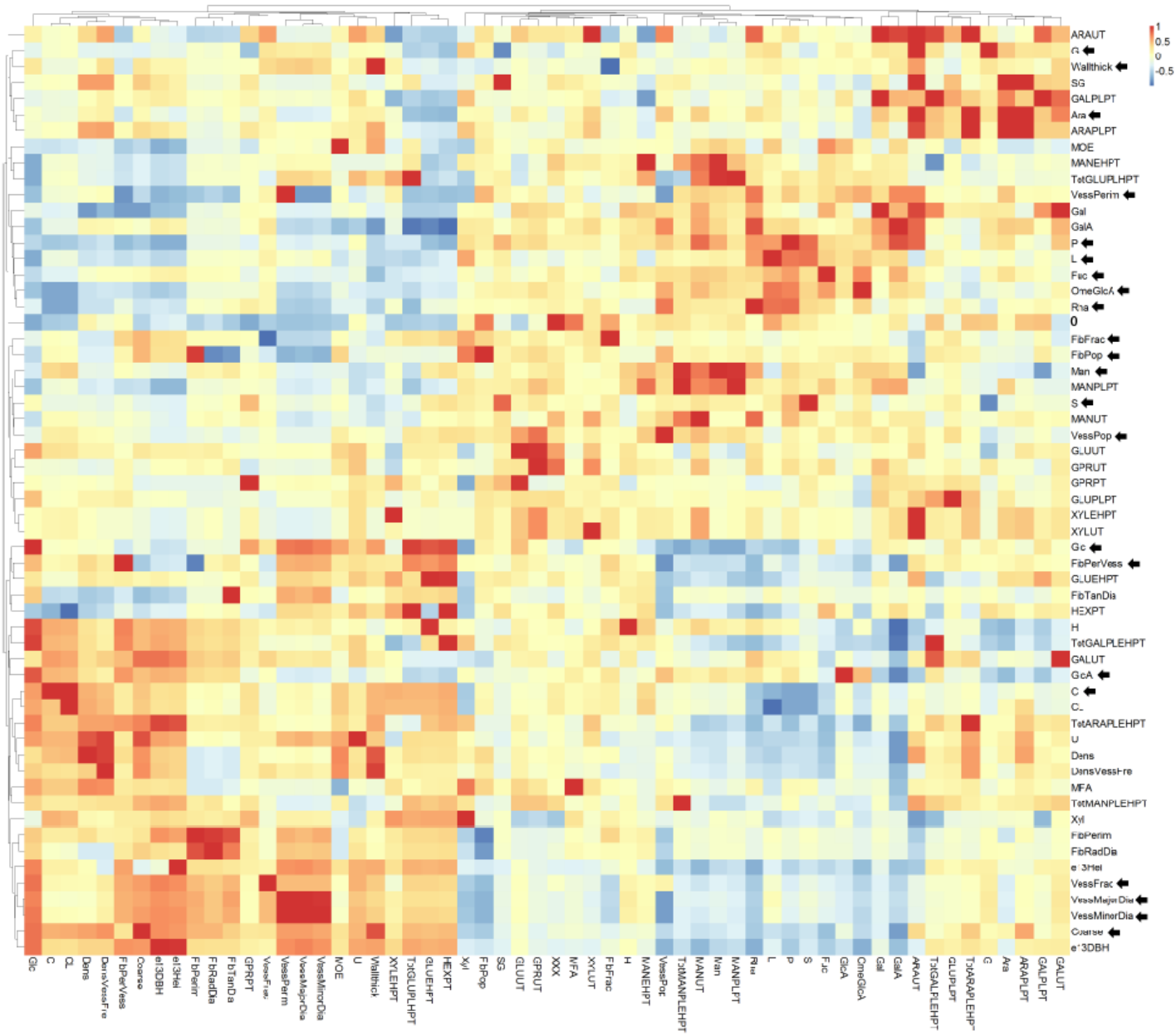

Figure S4: Genetic correlations between traits

Pairwise genetic correlations between traits with hierarchical clustering. Traits clustering together show similar genetic correlations with other traits, allowing to identify groups (clusters) of traits displaying the same pattern of how they genetically correlate to other traits. For clustering, the ”Ward method” was used as previously described: Murtagh, Fionn and Legendre, Pierre (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, **31**, 274--295. 10.1007/s00357-014-9161-z.

Traits marked by arrows are predictors of total-wood glucose yield.
Clarifications for traits’ names are supplied in Table S2.

Fig. S5

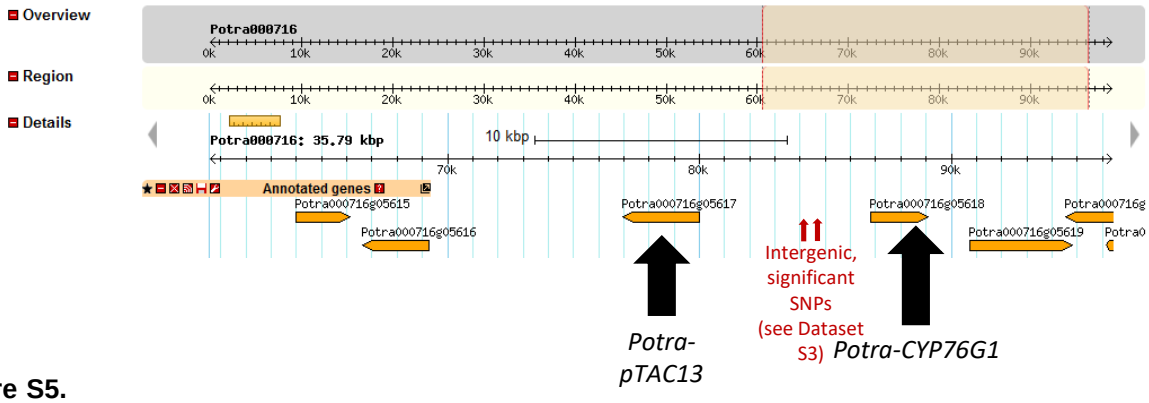

Figure S5.
Genome browser view of the genomic region where all the significant SNPs associated with G-lignin fall. 14 out of the 16 significant SNPs fall in and around the *Potra000716g05617/Potra-pTAC13* gene. The two other SNPs lay closest to the neighboring *Potra000716g05618/Potra-CYP76G1* gene, in the upstream sequence that may correspond to the distal part of the gene's promoter.

Fig. S6

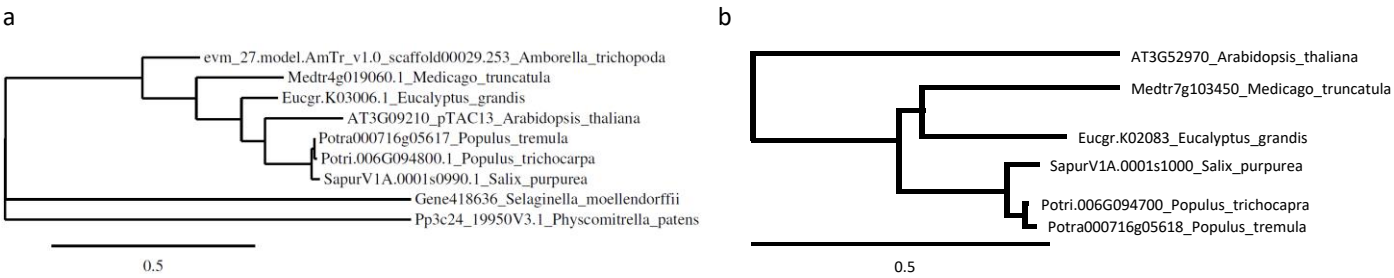

Figure S6.
Phylogenetic trees showing that the aspen G-lignin-associated genes (A) Potra000716g05617 and (B) Potra000716g05618 have an ortholog in other potential bioenergy feedstocks *P. trichocarpa*, *S. purpurea*, *E. grandis* and *M. truncatula*. Potra000716g05617 (but not Potra000716g05618) also has orthologs in more distantly related land plants such as Embryophyte *P. patens*, Tracheophyte *S. moellendorffii*, early Angiosperm *A. trichopoda*, and model dycotyledonous *Arabidopsis* (*A. thaliana*).