Supplementary Information on Materials and Methods

Cloning.

The cloning procedure used to generate already published constructs (Dataset S1) has been described in the corresponding publications (Dataset S1). In addition, the antisense constructs for the lines BI-20 and BI-22 were generated using a similar procedure as described for BI-23 (36).

For down-regulation lines using RNAi (Dataset S1), a collection of previously identified (37) hybrid aspen (*Populus tremula x tremuloides*) wood-expressed sequences (expressed sequence tags or ESTs) was used as a template to amplify the target sequences (as described in (38)). Gateway® cloning (Thermo Fisher Scientific, USA) was used to transfer each amplified sequence into the vector pK7GWiWG2(I) (38), thus generating a construct for RNAi down-regulation of the target gene.

For overexpression lines (Dataset S1), mRNAs were isolated from both leaves and stems of hybrid aspen (*Populus tremula* x *tremuloides* Michx.) T89 clones and the corresponding cDNA were synthesized. The cDNA of the target genes for overexpression were amplified and introduced into the overexpression vector pK2GW7 (39) using Gateway® cloning (Thermo Fisher Scientific, USA).

Plant material and growth conditions.

To generate the BioImprove collection, a subset of transgenic hybrid aspen (*Populus tremula* x *tremuloides* Michx.) lines was selected from previously characterized and published studies (Dataset S1). Most of the lines in the BioImprove collection were hybrid aspen (*Populus tremula* x *tremuloides* Michx.) T89 clones that were transformed with the aforementioned constructs using *Agrobacterium*-mediated gene transfer. Transformants were selected based on antibiotic resistance, grown and multiplied *in vitro* as previously described (33). For each construct, three to five different lines were tested in an earlier study for wood chemistry (40). From this study, we selected one transgenic line for each construct on the basis of giving the largest difference in wood chemistry. Fifty-one wild-type trees and four to five biological replicates for each transgenic line were grown for two months in previously described greenhouse conditions (17). Each tree's height, diameter (10 cm above ground) and mean internode length were measured, and 8-cm-long sections of stem were harvested 20 cm above ground. The bark was removed and the wood was freeze-dried and ground as previously described (17) to perform cell wall chemistry and saccharification analyses. The cut trees were allowed to re-grow one new shoot, which was repeatedly trimmed at the height of 1 meter. After 10 months (i.e. a total age of the plants of 12 months), an 8-cm-long piece of the main stem 10 cm above ground was collected, debarked, dried, and used to monitor the anatomical and structural features of the wood.

Saccharification.

As described previously (17), wood samples were freeze-dried and roughly ground. From the resulting powder, the fraction encompassing particle sizes from 0.1 mm to 0.5 mm was collected for further processing. For each sample, 50 mg dry weight of substrate were submitted (or not) to an acidic pretreatment (1% (w/w) sulphuric acid) during 10 min at 165 °C using a single-mode microwave system (Initiator Exp, Biotage, Sweden). The resulting samples were centrifuged 15 min at 14,100g in order to separate the solid fraction from the so called pretreatment liquid. The solid fraction from pre-treated samples was washed with deionized water and with sodium citrate buffer (see details in (17)). Both pretreated and non-preteated samples were submitted to enzymatic hydrolysis 72 h at 45 °C under agitation, using 25 mg of commercial enzymes mixture containing equal proportion (w/w) of Celluclast 1.5L and Novozyme 188 (Sigma-Aldrich, USA). The resulting liquid hydrolysates, as well as the above pretreatment liquid fractions, were analyzed using high-performance anion-exchange chromatography (HPAEC).

Wood anatomical and structural features.

SilviScan (CSIRO, Australia) measurements conducted at INNVENTIA were performed on all lines but three (BI-13, 21 and 26). Parallelepipedic radial pieces of wood were scanned with 2mm increments as described previously (35, 43, 44). The first measurement increment(s) covering not only wood but also the pith was (were) excluded from the analysis. Each remaining incremental measurement was weighted to reflect the total cross-sectional area that it represents in the wood. For each tree, the radial average was calculated for each trait measured by SilviScan (Dataset S2).

Mathematical modeling

Models were created for stem height, stem diameter, wood density and glucose release after pre-treatment and 72h enzymatic hydrolysis. To model these 4 traits based on wood biomass traits, these 4 traits were excluded from the set of traits used for modeling. In addition, the 19 remaining saccharification traits were also excluded from modeling because measuring any of them would allow, for technical reasons, to measure the others at the same time. Hence, glucose release would be measured at the same time as other saccharification traits, rendering its modeling superfluous.

Using R, numerous (≥30) models were generated with the aim of predicting each of the four traits used to calculate TWG (i.e. height, diameter, wood density and glucose release after pre-treatment). More precisely, for each of the four traits, three types of models were generated: (i) linear models which rely on linear relations between variables, (ii) Generalized Additive Models (GAMs (45, 46); package ”mgcv” (47)) which allow combining linear terms and different types of non-linear terms whose relations to the dependent variable can be represented by smooth functions, (iii) Random forests (48)(package ”Ranger” (49)) which rely on numerous tree predictors each using random subsets of independent variables in order to allow comparing the trees to reach an optimal prediction and to evaluate how much each variable contributes to this prediction. Numerous models from each type were generated for each trait by iteratively modifying parameters such as the input independent variables and the criteria for fitting (e.g. number of trees in Random forests or gamma for GAMs). Finally, the predictivity of each model was evaluated by calculating their Q^2, using a ”leave-one-out” approach. For each trait, the model from each type with the highest Q^2 value among its kind was selected (Dataset S4). Next, for each trait the type of model used *in fine* was also selected based on having the highest Q^2 compared with the other types of models (Dataset S4). Finally, the ultimately selected models for each trait were combined into a composite model to predict TWG and this composite model was evaluated for goodness of fit (R^2) and predictivity (Q^2).