Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing

Eshchar Mizrachi,a,b,1,2, Lieven Verbeke,c,d,1 Nanette Christie,a,b Ana C. Fierroc,d, Shawn D. Mansfield,⁎, Mark F. Davis,f, Erica Gjerding1, Gerald A. Tuskan9, Marc Van Montagu,d,h,2 Yves Van de Pee,b,d,h, Kathleen Marchalac,d,2, and Alexander Myburga,b,2

⁎Department of Genetics, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria 0028, South Africa; †Genomics Research Institute, University of Pretoria, Pretoria 0028, South Africa; ‡Department of Information Technology, Ghent University, iMinds, 9052 Gent, Belgium; §Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Gent, Belgium; †Department of Wood Science, University of British Columbia, Vancouver, BC, Canada V6T 1Z4; □Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401; †Department of Plant Systems Biology, VIB, 9052 Gent, Belgium

Contributed by Marc Van Montagu, December 12, 2016 (sent for review October 7, 2016; reviewed by Kenneth Keegstra and Joachim Selbig)

As a consequence of their remarkable adaptability, fast growth, and superior wood properties, eucalypt tree plantations have emerged as key renewable feedstocks (over 20 million ha globally) for the production of pulp, paper, bioenergy, and other lignocellulosic products. However, most biomass properties such as growth, wood density, and wood chemistry are complex traits that are hard to improve in long-lived perennials. Systems genetics, a process of harnessing multiple levels of component trait information (e.g., transcript, protein, and metabolite variation) in populations that vary in complex traits, has proven effective for dissecting the genetics and biology of such traits. We have applied a network-based data integration (NBDI) method for a systems-level analysis of genes, processes and pathways underlying biomass and bioenergy-related traits using a segregating Eucalyptus hybrid population. We show that the integrative approach can link biologically meaningful sets of genes to complex traits and at the same time reveal the molecular basis of trait variation. Gene sets identified for related woody biomass traits were found to share regulatory loci, cluster in network neighborhoods, and exhibit enrichment for molecular functions such as xylan metabolism and cell wall development. These findings offer a framework for identifying the molecular underpinnings of complex biomass and bioprocessing-related traits. A more thorough understanding of the molecular basis of plant biomass traits should provide additional opportunities for the establishment of a sustainable bio-based economy.

significance

Carbon fixation and accumulation as lignocellulosic biomass is of global ecological and industrial importance and most significantly occurs in the form of wood development in trees. Traits of importance in biomass accumulation are highly complex and, aside from environmental factors, are affected by many pathways and thousands of genes. We have applied a network-based data integration method for a systems genetics analysis of genes, processes, and pathways underlying biomass and bioenergy-related traits using segregating Eucalyptus hybrid tree populations. We could link biologically meaningful sets of genes to complex traits and at the same time reveal the molecular basis of trait variation. Such a holistic view of the biology of wood formation will contribute to genetic improvement and engineering of plant biomass.


Reviewers: K.K., Michigan State University; and J.S., Max Planck Institute of Molecular Plant Physiology.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the NCBI Sequence Read Archive (accession no. SUB2087452).

1E.M. and L.V. contributed equally to this work.

2To whom correspondence may be addressed. Email: mamoni@pab.ugent.be, eshchar.mizrachi@fab.iap.ac.za, kathleen.marshall@intech.ugent.be, or zander.myburg@fab.iap.ac.za.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1620119114/-/DCSupplemental.
study woody biomass traits in long-lived perennial plants. While such association mapping approaches facilitate the delimitation of genome positions harboring causal variation (13), they provide little information as to how these genes and their variants act together in biological pathways to influence trait variation.

Complementing genetic information with molecular phenotypes (e.g., transcript levels) can contribute to a better mechanistic understanding of trait variation (24, 26, 27). Expression QTL (eQTL) analysis (28) allows the identification of genomic loci associated with variation in molecular phenotypes. In contrast to QTL or association analyses, eQTL analysis also identifies the genes that are affected by this variation and thus potentially contribute to the complex trait. However, because complex traits are subject to the combined effect of multiple genetic loci, each with a small effect on the trait, there is generally low power to detect statistically significant associations when relying on single-gene analysis. In addition, single gene associations do not show how genes and/or pathways interact to explain a complex trait (29).

To cope with the aforementioned limitations of the association problem, more integrated systems genetics approaches have been proposed (29, 30). Methods that use network models to represent molecular a priori knowledge on the organism/trait of interest (31) have been particularly successful to perform association analysis in clonal systems (32–35). In the context of outbred populations, network-based methods have been applied for gene prioritization (36) or to increase the reliability of eQTL association mapping itself (37) but not yet for integrative association analysis.

Here we applied a systems genetics approach to study the genomic loci and pathways affecting wood formation in *Eucalyptus*. We generated coupled genetics/genomics (linkage map and immature xylem transcriptome) data for 156 individuals segregating from an F2 pseudobackcross between a *Eucalyptus grandis* × *Eucalyptus urophylla* F1 interspecific hybrid tree and an unrelated *E. urophylla* tree and profiled traits representative of tree growth (diameter at breast height and bark thickness), wood properties (wood basic density and cell wall composition), and bioprocessing metrics (sugar release). Data were integrated using a unique network-based data integration (NBDI) approach that allows combining genotyping, expression profiling, and prior network information to prioritize genes and molecular mechanisms associated with complex wood formation traits.

### Table 1. Comparison of using NBDI-transformed versus using nontransformed expression values in prioritizing reference genes for 13 traits

<table>
<thead>
<tr>
<th>Trait</th>
<th>NBDI enrichment $p$</th>
<th>Nontransformed enrichment $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBH (over bark)</td>
<td>$6.38 \times 10^{-20}$</td>
<td>$2.81 \times 10^{-11}$</td>
</tr>
<tr>
<td>DBH (under bark)</td>
<td>$5.55 \times 10^{-15}$</td>
<td>$1.38 \times 10^{-10}$</td>
</tr>
<tr>
<td>Bark thickness</td>
<td>$4.61 \times 10^{-19}$</td>
<td>$1.38 \times 10^{-10}$</td>
</tr>
<tr>
<td>Wood density</td>
<td>$6.52 \times 10^{-10}$</td>
<td>$5.49 \times 10^{-12}$</td>
</tr>
<tr>
<td>Lignin content</td>
<td>$2.83 \times 10^{-10}$</td>
<td>$9.28 \times 10^{-10}$</td>
</tr>
<tr>
<td>Total C6 sugar in walls</td>
<td>$1.02 \times 10^{-01}$</td>
<td>$2.83 \times 10^{-01}$</td>
</tr>
<tr>
<td>Total C5 sugar in walls</td>
<td>$1.02 \times 10^{-01}$</td>
<td>$9.75 \times 10^{-01}$</td>
</tr>
<tr>
<td>Glucose released</td>
<td>$5.49 \times 10^{-12}$</td>
<td>$3.03 \times 10^{-04}$</td>
</tr>
<tr>
<td>Percent of max glucose</td>
<td>$3.29 \times 10^{-14}$</td>
<td>$1.03 \times 10^{-04}$</td>
</tr>
<tr>
<td>Xylose released</td>
<td>$1.88 \times 10^{-13}$</td>
<td>$1.02 \times 10^{-01}$</td>
</tr>
<tr>
<td>% of max xylose release</td>
<td>$2.21 \times 10^{-03}$</td>
<td>$7.19 \times 10^{-01}$</td>
</tr>
<tr>
<td>Glucose + xylose released</td>
<td>$2.81 \times 10^{-11}$</td>
<td>$3.03 \times 10^{-04}$</td>
</tr>
<tr>
<td>Percent of max sugar released</td>
<td>$3.29 \times 10^{-14}$</td>
<td>$2.21 \times 10^{-03}$</td>
</tr>
</tbody>
</table>

For each trait, 300 genes were selected. The enrichment score corresponds to the $P$ value of a hypergeometric enrichment test.

### Results

**Network-Based Gene–Trait Association.** One hundred fifty-six *E. grandis* × *E. urophylla* F2 interspecific backcross trees were profiled for transcript abundance in immature xylem and for woody biomass traits that relate to growth, wood density, cell wall composition, and sugar extractability (Table 1 and Materials and Methods). Genotyping, expression profiling, inferred eQTL associations, and prior network information were simultaneously used to prioritize genes and molecular mechanisms associated with each of the measured traits. To this end we developed an integration approach that makes use of a gene interaction network model in which nodes are genes and edges represent two types of information. If derived from prior information, edges reflect relations between genes and gene products, derived from Kyoto Encyclopedia of Genes and Genomes (KEGG) (SI Materials and Methods). If derived from eQTL associations, edges reflect that the connected genes share the same eQTL and thus are likely functionally related and coregulated.

This network model is then used to propagate on a per sample basis the expression signals of genes to a local network neighborhood. This propagation transforms the original gene expression data to network-diffused gene expression data (Fig. 1 and Materials and Methods). In the network-diffused expression matrix, each data point can be interpreted as the original expression signal of a gene in a sample, modulated by the expression of the genes that are close neighbors in the network (i.e., that are likely found in the same pathways or to share eQTLs, etc.). Modulation implies that if nodes in the local neighborhood of a gene are
Genes and Pathways of Relevance to the Traits Under Study. The NBDI approach combines genetic and prior information with gene expression variation to prioritize, per trait, relevant genes/pathways influenced by genetic variation in the population. This combined information is captured in the two complementary views, an eQTL (Fig. 3) and a network view (Fig. 4).

First, because of its network model, NBDI implicitly imposes that genes prioritized for a trait should also share an eQTL at one or more loci in the genome. Most eQTLs for genes prioritized by the NBDI approach for a particular trait should therefore cluster together in hot spots rather than being randomly scattered along the genome. Fig. 3 shows this is indeed the case.

Second, as the NBDI approach favors genes that are connected in the interaction network, at least some of the genes associated with a trait, when projected on a gene interaction network, should cluster together and constitute a molecular subnetwork underlying the trait providing a complementary network view. Because this network of curated gene interactions (derived from KEGG) is sparser than the network used for the NBDI analysis that also includes eQTL overlap relations, by definition only a subset of the gene selections and relations can be visualized in subnetworks. To visualize eQTL relations for genes that could be connected through KEGG, eQTL overlap relations were overlaid on the identified subnetworks. For each trait, the largest connected component of the obtained network is

![Fig. 2. Performance comparison of using NBDI-transformed (blue starred line) versus using nontransformed expression correlation data (red line) in prioritizing trait-associated genes.](image)

![Fig. 3. Location of eQTLs shared by the genes selected for each individual trait.](image)

Mizrachi et al. PNAS Early Edition | 3 of 6
extracted, containing the prioritized genes that can be connected in the network through direct edges. Representative subnetworks of the broad trait classes (Fig. 3) are presented in Fig. 4. The networks corresponding to wood density, diameter at breast height (DBH), glucose released, and lignin content contain 20, 39, 28, and 48 genes, respectively, from the original 300 selected genes per trait and were highly significant (the probability of obtaining a connected component with the observed size purely by chance was smaller than $10^{-10}$, $10^{-23}$, $10^{-17}$, and $10^{-32}$, respectively).

As shown in Figs. 3 and 4, the traits under study depend, at least partially, on similar processes and shared eQTLs. This is expected given that all traits reflect wood-related properties and will be, to different extents, phenotypically related. However, trait-specific differences can be identified. When the traits are clustered based on the eQTLs shared by the genes in their gene selections (SI Materials and Methods), four broad groups of traits can be identified (Fig. 3): (1) wood density, (2) sugar release (bioprocessing) metrics, (3) growth traits, and (4) C5 and C6 sugar in cell walls together with lignin content and the percent of maximum xylose release. Below, the biological functions of the genes associated with the shared eQTL peaks found for traits and trait groups are discussed in more detail (Fig. S4 and Dataset S3).

The biological functions of the genes associated with traits in groups 2 (sugar release) and 3 (growth traits) are characterized by shared cell wall related processes (cell wall organization and bioprocessing, hemicellulose metabolic processes, xylan biosynthesis, glucuronoxylan biosynthesis, and lignin biosynthesis) involving genes with eQTL located at peaks 43, 48, 49, and 50, and by processes that are shared by all traits at eQTL peaks 10 and 11 (but for which no functional overrepresentation could be assigned) (Fig. S4). All growth related traits (bark thickness and DBH) and all bioprocessing (sugar release) metrics (except percent maximum xylose release) belong to this group.

Growth-related traits (group 3) seem to be dominated, in addition to the cell wall related processes mentioned above, by anthocyanin related processes (eQTL peak 33). This is also illustrated in the DBH subnetwork that is representative for this trait group (Fig. 4). The DBH subnetwork contains a considerable number of lignin-related genes. Most bona fide lignin biosynthesis genes that are highly expressed in developing xylem (8, 38) (Dataset S3) were indeed associated with growth (DBH and bark thickness traits). In addition to these lignin-related genes, the DBH network contains several genes involved in hormone signaling (IAA9 and ARF7) and flavonoid biosynthesis (ANS, BAN, and A11). The relationship between lignin and flavonoid biosynthesis has been shown in Arabidopsis, where silencing of hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase (HCT) resulted in decreased plant growth and redirection of the metabolic flux into flavonoid production through chalcone synthase (39).

In contrast, sugar release traits (trait group 2) and wood density (trait group 1) differ from the growth-related traits (trait group 3) in their larger involvement of photosynthesis-related processes (related to eQTL peak 46; Dataset S3). Because secondary xylem tissue is the main carbohydrate sink in woody plants, its composition is expected to be affected by the availability of fixed carbon resulting from photosynthesis (40). A representative subnetwork of this group of sugar release traits (percent glucose released; Fig. 4) is much less dominated by prior interactions (they might be less known for the processes involved) but clearly shows the presence of two clusters of genetic associations involving, among others, genes encoding TUBULIN ALPHA-2 CHAIN (TUA2), COBALAMIN-INDEPENDENT METHIONINE SYNTHASE/METHIONINE SYNTHESIS 1 (ATCMS/MS1), and IRREGULAR XYLEM 8/GALACTURONOSYLTRANSFERSASE 12 (IRX8/GAUT12), homologs of which have previously been identified as associated with hemicellulose and lignin related properties in Populus (24). IRX8/GAUT12 is known to be involved in xylan structure (41), and modulation of its expression in poplar results in improved sugar release efficiency (42). More indirect arguments can be made for TUA2 and ATCMS/MS1. Tubulins affect cortical microtubule arrangement and thus cellulose microfibril angle (MFA) (43, 44). TUA2 specifically is known to be a target of SND1, the master regulator of secondary cell wall deposition (45, 46), and is one of two alpha-tubulins that is significantly up-regulated in Eucalyptus during tension wood formation, during which MFA is one of the main physical changes in the wood (47). Given that MFA is thought to affect wood ultrastructure and stiffness (48, 49), it is interesting to find this tubulin associated with bioprocessing-related traits such as sugar release in this study. Several associations were identified (mainly with sugar release efficiency) for genes involved in cysteine and methionine metabolism, including ATCMS/MS1. The roles of these genes in biomass formation are becoming increasingly revealed, being linked directly to either lignin (50, 51) or hormone-mediated growth regulation (52). In addition, SCW polysaccharide biosynthesis genes known to be expressed in developing xylem (8) were mainly associated with variation in wood density and glucose release efficiency (Dataset S3). In the latter case, the majority were xylan modification genes, affecting patterns of acetylation, and glucuronic acid and methyl-glucuronic acid decoration of the xylan backbone (53, 54) (Dataset S3).
Trait group 4 contains traits related to the total cell wall sugar content and, surprisingly, also lignin content and percent maximum xylose released. These traits are characterized by the relatively smaller effect of major eQTL peaks that dominate most of the other biomass traits. The lignin subnetwork (Fig. 4) in general lacks most of the genes related to lignin biosynthesis itself. Indeed, few association between the variation of expression of SCW biosynthetic genes (cellulose, xylan, and lignin pathway genes) and the final C5 and C6 sugar and lignin content of the cell wall were apparent (Dataset S3). Several genes highly associated with variation in lignin content code for enzymes involved in carbon metabolism, including phosphopentofructose (phosphopentofructose carboxylkinase enolase), pyruvate (plastidial pyruvate kinase 3 and malate dehydrogenase), and acetyl-CoA (pyruvate dehydrogenase E1 α subunit) metabolism, as well as pathways producing UDP-glucose and fructose (UDP-glucose pyrophosphorylase and sucrose synthase; Dataset S3). Genes involved in mitochondrial energy metabolism were also associated with lignin content. Very distinctive for this group of traits are also the genes that relate to abiotic stress (eQTL peak 34) and RNA modification (a process associated with a very distinctive eQTL peak 48).

As an additional external validation, we overlaid previously identified QTLs (blue circles in Fig. 3) for the same complex traits (20) with the obtained eQTL frequency peaks (Fig. 3). These results show that at least some of these trait QTLs are in close proximity to eQTL frequency peaks (especially peaks 28, 29, 48, and 49, located on chromosomes 6 and 10), providing additional evidence that the gene selection is relevant for the trait under study. Several eQTL peaks cannot be directly mapped to trait QTLs. These might represent polymorphisms that only have detectable effects on molecular subcomponents of a trait but cannot be directly associated with the phenotype itself. Given that complex traits are affected by different molecular traits in epistatic and nonlinear ways, a direct link between molecular traits and phenotype traits is not always expected or the effects are too numerous and small to detect at the level of trait QTLs given the relatively small size of experimental population (n = 156).

**Discussion**

The observed variability of woody biomass traits in this study is explained by the variation of combinations of genes or sets of closely interacting pairs of genes, each influenced by genetic variation segregating in this particular interspecific backcross population. Because of this, linking a quantitative trait to the expression of individual genes might fail or be incomplete if the trait under study is influenced by variation in the expression of large numbers of genes that in turn can be influenced by the expression of other genes, etc. If this is the case, then any method that captures only the marginal effect of a gene on a trait might render only a partial view of the genes that are involved in the underlying biological processes. To cope with this statistical issue we have developed a network-based data integration approach (NBDI) that combines genotyping, expression profiling, and prior network information to prioritize genes and molecular mechanisms associated with measured traits.

This NBDI approach is based on a network model in which connections between genes reflect interactions derived from either prior molecular interaction information or from eQTL information. In the latter case it is assumed that if two genes share an eQTL, they are connected in the network because of a shared coregulation mechanism. Even though incidental overlap of eQTLs is possible, for instance, through the action of separate polymorphisms in tightly linked but unrelated genes, we assumed that the majority of the overlapping trans-eQTLs can be treated as evidence of a shared regulatory polymorphism, as reflected by the shared functional annotations observed for the associated genes. Gene expression signals are then propagated through the network model to obtain an integrated signal that is used to explain the variation in the external traits.

We applied the NBDI approach to study the genomic loci and pathways affecting wood formation in *Eucalyptus*. The experimental setup used [with high linkage disequilibrium (LD) and large effect QTLs segregating in a single family] is complementary to low-LD studies (with high resolution but typically small effect associations) in populations of unrelated individuals (e.g., refs. 24, 55).

Using our integrative systems genetics approach allowed for prioritizing genes contributing to woody biomass traits and identifying the putative regulatory loci with which these genes and traits are predominantly associated. Based on this analysis, a clear distinction could be made between growth and sugar release (bioprocessing) related traits and traits related to the total cell wall sugar content. Unexpectedly, we noticed little association between the variation of expression of SCW biosynthetic genes (cellulose, xylan, and lignin pathway genes) and the final C5 and C6 sugar and lignin content of the cell wall. Rather, most bona fide lignin biosynthesis genes were associated with growth-related traits (DBH and bark thickness), and most SCW polysaccharide biosynthesis genes were associated with variation in wood density and glucose release efficiency. Several genes highly associated with variation in lignin content code for enzymes involved in carbon metabolism and in mitochondrial energy metabolism. As a result, we hypothesize that variation in the expression of SCW biosynthetic genes has an effect on the growth and ultrastructure and resultant processability of the secondary cell wall, whereas the quantity of sequestered carbon in the cell wall (in the form of polysaccharides and phenolics) is more related to variation in primary carbon metabolism pathways and hence precursor availability. This assumption further establishes the strong link between physiological/cellular homeostasis and secondary processes such as SCW polysaccharide and lignin biosynthesis that represent a strong, irreversible carbon sink in woody plants.

**Materials and Methods**

**Experimental Population, Transcriptome, and Complex Trait Analysis.** The F2 backcross population was generated from a cross between an *E. grandis* × *E. urophylla* F1 interspecific hybrid parent (GUSAP1, Sappi Forest Research, South Africa) and an unrelated *E. urophylla* parent (USAP1) (18). At 3 y old, immature xylem tissue was harvested from 156 individuals as previously described (56). Samples were collected from 3-y-old trees over a 7.5-h period between 0900 and 1630 hours for 3 d. Total RNA was isolated (57) and used for RNA-Seq expression profiling (30 million; Illumina PE50, BG Hong Kong). Gene expression values (FPKM) were calculated per gene model using TopHat version 1.3 and Cufflinks version 1.0.3 (bias correction and quartile normalization was enabled for the FPKM calculation) (58, 59). Diameter (cm) at breast height (DBH) of the main stem was assessed as described previously (20). Bark thickness was calculated as the difference between over-bark and under-bark DBH measurements. A wood disk taken at breast height (1.35 m) was used to determine wood basic density using the water displacement method (www.tappi.org/content/SARG/T258.pdf). Chemical wood properties were assessed using different analytical methods, including pyrolysis molecular beam mass spectrometry (pyMBMS).

Trait QTL mapping, eQTL mapping, and eQTL classification are described in SI Materials and Methods.

**NBDI Association Analysis.** First, a hybrid gene interaction network was constructed using curated gene interactions downloaded from KEGG and eQTL overlap relations (Fig. 1A). For the latter, we investigated for pairs whether these genes had overlapping eQTL intervals (Fig. S5). If this is the case, a connection in the hybrid network is added. Once the network is constructed, a graph node kernel was calculated (the Laplacian exponential diffusion kernel; 60) to quantify how well each node in the network is connected to other nodes (Fig. 1B). The transformed connectivity matrix was then multiplied with the gene expression matrix to obtain the diffused or transformed gene expression matrix (Fig. 1C). Genes in the network connectivity matrix that were not present in the gene expression matrix were removed and vice versa. The transformed gene expression was finally linked to the measured traits by calculating the absolute value of the Pearson correlation between the transformed gene expression and the measured traits. After ranking, the top 300 genes exhibiting the highest correlation were selected for further analysis. For details of the eQTL overlap procedure, network construction, connectivity calculation, and association analysis, see SI Materials and Methods and Figs. S5 and S6.
ACKNOWLEDGMENTS. This work was supported by the Department of Science and Technology (Strategic Grant for the Euclayptus Genomics Platform) and National Research Foundation of South Africa (Bioinformatics and Functional Genomics Programme, Grants 86936 and 97911 to A.A.M.), Sappi South Africa and the Technology and Human Resources for Industry Programme (Growthpoint) through the Forest Molecular Genetics Programme at the University of Pretoria (to A.A.M.), Ghent University Multi-disciplinary Research Partnership from nucleotides to networks (Project 01MR0410W to V.V.D. and K.M.), the European Union (FP7/2007-2013) through the project EUROBRIDGE (Grant 322739) and the U.S. Department of Energy Office of Science (G.A.T.). Finally, the authors acknowledge Sappi Forest Research for the plant materials and growth and wood property data used in the study.