Functional investigation of five R2R3-MYB transcription factors associated with wood development in *Eucalyptus* using DAP-seq-ML

Lazarus T. Takawira[1]*, Ines Hadj Bachir[2]*, Raphael Ployet[1], Jade Tulloch[1], Helene San Clemente[2], Nanette Christie[1], Nathalie Ladouce[2], Annabelle Dupas[2], Avanish Rai[1], Jacqueline Grima-Pettenati[2], Alexander A. Myburg[1], Eshchar Mizrachi[1], Fabien Mounet[2†], Steven G. Hussey[1‡]

*These authors contributed equally to the study

[†]Corresponding author: fabien.mounet@univ-tlse3.fr

[‡]Corresponding author: steven.hussey@fabi.up.ac.za

[1]Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria 0002, South Africa

[2]Laboratoire de Recherche en Sciences Végétales, Université Toulouse, CNRS, INP, Castanet-Tolosan, France

Key message

We combined DAP-seq with machine learning to identify biologically relevant gene targets for five *Eucalyptus* R2R3-MYB transcription factors linked to wood formation, additionally demonstrating that EgrMYB137 promotes lignification *in planta*.

Abstract

A multi-tiered transcriptional network regulates xylem differentiation and secondary cell wall (SCW) formation in plants, with evidence of both conserved and lineage-specific SCW network architecture. We aimed to elucidate the roles of selected R2R3-MYB transcription factors (TFs) linked to *Eucalyptus* wood formation by identifying genome-wide TF binding sites and direct target genes through an improved DAP-seq protocol combined with machine learning for target gene assignment (DAP-seq-ML). We applied this to five TFs including a well-studied SCW master regulator (EgrMYB2; homolog of AtMYB83), a repressor of lignification (EgrMYB1; homolog of AtMYB4), a TF affecting SCW thickness and vessel density (EgrMYB137; homolog of PtrMYB074) and two TFs with unclear roles in SCW regulation (EgrMYB135 and EgrMYB122). Each DAP-seq TF peak set (average 12,613 peaks) was enriched for canonical R2R3-MYB binding motifs. To improve the reliability of target gene assignment to peaks, a random forest classifier was developed from *Arabidopsis* DAP-seq, RNA-seq, chromatin, and conserved noncoding sequence data which demonstrated significantly higher precision and recall to the baseline method of assigning genes to proximal peaks. EgrMYB1, EgrMYB2 and EgrMYB137 predicted targets showed clear enrichment for SCW-related biological processes. As validation, *EgrMYB137* overexpression in transgenic *Eucalyptus* hairy roots increased xylem lignification, while its dominant repression in transgenic *Arabidopsis* and *Populus* reduced xylem lignification, stunted growth, and caused downregulation of SCW genes. EgrMYB137 targets overlapped significantly with those of EgrMYB2, suggesting partial functional redundancy. Our results show that DAP-seq-ML

identified biologically relevant R2R3-MYB targets supported by the finding that EgrMYB137 promotes SCW lignification *in planta*.

Keywords: DAP-seq; machine learning; secondary cell wall; R2R3-MYB; transcription factors; *Eucalyptus*

Introduction

Plants are estimated to constitute ~550 billion tons (~80%) of the world's carbon biomass (Bar-On et al. 2018), the majority of which is comprised of lignified secondary cell walls (SCWs). The deposition of lignin, cellulose, and hemicellulose in SCWs during wood formation depends on the highly coordinated regulation of SCW structural genes and transcription factors (TFs) (Kumar et al. 2016; Houston et al. 2016; Sundell et al. 2017; Luo and Li 2022). A semi-hierarchical transcriptional network predominantly featuring NAC and R2R3-MYB families regulates SCW development (Zhong and Ye 2007; Zhong and Ye 2015; Zhang et al. 2018; Hussey 2022). The lower tier predominantly regulates SCW structural genes while R2R3-MYB TFs occupy the central position in the network, acting as hub regulators of SCW deposition. These are in turn activated by NAC domain master regulators, while the NAC master regulators are themselves under the control of HD-ZIP III, R2R3-MYB, and AS/LBD TFs which are activated by auxin response factors (Qu et al. 2021).

The R2R3-MYB family is one of the largest TF groups in higher plants and regulates diverse biological processes (Jin and Martin 1999; Zhong et al. 2008; Wang et al.

2009; Dubos et al. 2010; Ambawat et al. 2013; Vimolmangkang et al. 2013; Roy 2016) . Several MYB-family TFs such as AtMYB4, AtMYB43, AtMYB46, AtMYB61 and AtMYB83 have been identified as regulators of SCW development (Zhong et al. 2008; Wang et al. 2009; Dubos et al. 2010; Ambawat et al. 2013; Vimolmangkang et al. 2013; Nakano et al. 2015; Roy 2016; Xiao et al. 2021) . R2R3-MYB TFs recognise canonical AC-I (ACCTACC), AC-II (ACCAACC), and AC-III (ACCTAAC) motifs that are conserved across diverse angiosperm and gymnosperm lineages (Hatton et al. 1995; Raes et al. 2003; Patzlaff et al. 2003; Zhong et al. 2008; Zhou et al. 2009; Zhong and Ye 2012) .

There is strong evidence that the *Arabidopsis thaliana* paralogs AtMYB46 and AtMYB83, as well as their homologs in other plants, redundantly regulate the entire SCW program including lignification and programmed cell death (Goicoechea et al. 2005; Zhong et al. 2008; McCarthy et al. 2009; Zhong et al. 2010; Zhong and Ye 2012; Zhong et al. 2013; Ko et al. 2014). EguMYB2, a close homolog of AtMY46/AtMYB83 in *Eucalyptus gunnii*, increased lignification and SCW thickening when overexpressed in tobacco and activates the promoters of two known *Eucalyptus* lignin biosynthetic genes cinnamyl alcohol dehydrogenase (CAD) and cinnamoyl-coenzyme A reductase (CCR) (Goicoechea et al. 2005). The genome-wide targets of EguMYB2 are unknown, but those of its homologs have been revealed in *Arabidopsis* and *Populus* (Ko et al. 2009; Zhong and Ye 2012; Chen et al. 2019) .

Some R2R3-MYB TFs in the SCW transcriptional network carry out repressive functions. Studies in *Arabidopsis*, *Pinus*, *Panicum*, *Eucalyptus*, *Populus*, and *Zea*

found that AtMYB4 and its homologs, which contain a C-terminal repression domain, repress the lignin biosynthesis pathway (Jin et al. 2000; Patzlaff et al. 2003; Fornalé et al. 2010; Legay et al. 2010; Shen et al. 2012; Wang et al. 2020a) . AtMYB4-mediated repression of cinnamate 4-hydroxylase (C4H) leads to the downregulation of hydroxycinnamic acid biosynthesis (Jin et al. 2000). Additionally, environmental factors such as UV-B light exposure or wounding causes a reduction in *AtMYB4* expression leading to the de-repression of *C4H*, in the process achieving increased UV-protectant sinapate ester production (Jin et al. 2000). In switchgrass (*Panicum virgatum*), PvMYB4, a homolog of AtMYB4, downregulates monolignol pathway genes and its ectopic overexpression resulted in reduced lignin content and increased sugar release efficiency from cell wall residues (Shen et al. 2012). In *E. gunnii*, EguMYB1 interacts with the linker histone variant EguH1.3, together inhibiting lignin deposition in xylem cells through the repression of EguMYB1 targets. Through this mechanism, it is proposed that chromatin-regulating factors interact with transcriptional regulators to fine-tune lignin deposition (Soler et al. 2017) . Although the biological roles of EguMYB1 and EguMYB2 have been investigated, knowledge of their genome-wide target genes are lacking.

AtMYB52 is a direct target of AtMYB46/83 (Ko et al. 2009) and possibly occupies the lowest tier of the SCW network (Zhong et al. 2008; Ko et al. 2009; Zhou et al. 2009; Romano et al. 2012; Zhang et al. 2018) . Although there is limited functional genetics data, dominant repression of AtMYB52 resulted in a substantial decrease in fibre SCW deposition in *Arabidopsis* inflorescence stems and hypocotyls (Zhong et al. 2008; Ko et al. 2009; Taylor-Teeples et al. 2015) . *AtMYB52*, which is co-expressed

with many SCW-related genes, is associated with increased and ectopic lignification in the loss-of-function mutant (Cassan-Wang et al. 2013). There is also strong evidence linking it to the regulation of pectin formation: acting downstream of BLH2/BLH4, AtMYB52 inhibits homogalacturonan de-methylesterification in the seed coat mucilage by directly repressing several pectin methylesterases (PMEs). Additionally, the AP2/ERF family TF AtERF4 antagonises AtMYB52 through physical interaction at the same gene promoters to promote PME gene expression (Shi et al. 2018; Xu et al. 2020; Ding et al. 2021).

The SCW-related TF AtMYB43 together with AtMYB20, AtMYB42, and AtMYB85 directly activate phenylpropanoid biosynthesis genes and lignin biosynthesis genes during SCW development (Zhong et al. 2008; Zhong and Ye 2012; Geng et al. 2020). When these TFs were disrupted, reduced lignin content was observed in *Arabidopsis* (Zhong et al. 2008; Geng et al. 2020). In addition, the AtMYB4 repressor is activated by these four TFs, in the process optimizing phenylpropanoid flux to the lignin biosynthesis pathway through the repression of flavonoid biosynthesis (Geng et al. 2020). Thus, AtMYB43 and AtMYB52 appear to be implicated in SCW-related biological processes but knowledge of their precise target genes is still incomplete. Homologs of *AtMYB43* and *AtMYB52* in *Eucalyptus*, *EgrMYB122* and *EgrMYB135*, are preferentially expressed in xylem and considered to be largely uncharacterized regulators of wood tissues formation (Soler et al. 2016).

Nakano et al. (2015) highlighted the central roles of NAC and MYB family TFs in SCW regulation during the evolution of regulatory systems that aided SCWs to adapt to the

need for water conductivity and mechanical support of plants on land. The evolution and functional analysis of these two key TF families have also been studied in *Eucalyptus* (Soler et al. 2015; Hussey et al. 2015). Recent studies have demonstrated that *Populus* MYB TFs with no close homologs in *Arabidopsis* are involved in SCW biogenesis, and that many targets of even conserved regulators appear to be novel. Among them are PtrMYB074 (closest homolog of EgrMYB137) and PtrMYB021 (a homolog of AtMYB46/83) which are transcriptional activators of vessel and fibres SCW development (Chen et al. 2019; Ployet et al. 2019; Wang et al. 2020b; Liu et al. 2021). PtrMYB074, which interacts with both PtrMYB021 and PtrWRKY19, directly activates a host of SCW-related TFs in poplar, while one of its direct targets, PtrbHLH186, regulates vessel size and density (Chen et al. 2019; Liu et al. 2022). A close homolog of PtrMYB074, EgrMYB137, is impacted by water and/or K-fertilization limitation and potentially plays a role in SCW remodelling in stress conditions (Ployet et al. 2019). These findings support that PtrMYB074, EgrMYB137 and other homologs in woody dicots have roles that integrate wood formation and abiotic stress modules.

A mechanistic understanding of SCW-related transcriptional regulatory networks in *Eucalyptus* and other woody species requires genome-wide information on TF gene targets in different contexts. Physical interaction data is typically derived from approaches such as ChIP-seq and assigning putative target genes to *in vivo* TF binding sites (TFBS). However, ChIP-seq is low throughput in that quality custom antibodies are required for each TF candidate. Alternatively, *in vitro* approaches can identify TFBSs using DNA Affinity Purification sequencing (DAP-seq) (Bartlett et al.

2017). The method uses *in vitro*-expressed TFs to identify binding sites on chromatin-free methylated or unmethylated genomic DNA, allowing for high-throughput genome-wide TFBS identification (O'Malley et al. 2016; Bartlett et al. 2017). In most studies using DAP-seq to identify target genes (Galli et al. 2018; Lai et al. 2020; Zhu et al. 2021; Han et al. 2021; da Silveira Falavigna et al. 2021), gene targets were assigned to TFBSs through proximity-based methods. These are prone to high false positive rates, as is typically seen in ChIP-seq studies (e.g., Kaufmann et al. 2009). Furthermore, since DAP-seq does not account for intrinsic factors like native chromatin state or co-factors which affect TFBSs *in vivo*, many DAP-seq TFBSs are presumably *in vitro* anomalies. To date, there have been few efforts made to identify biologically relevant binding sites from DAP-seq data, optimise DAP-seq data quality, assess and develop benchmarks for its reproducibility, or improve methods for target gene assignment to arrive at high-confidence peak sets and target genes. Supervised approaches that incorporate complementary information on TF binding and gene regulation such as chromatin state and gene expression data might aid in the identification of *bona fide* target genes when trained on an exhaustive set of known, labelled TF-target associations and features predictive of these associations.

In this study, we used DAP-seq to reconstruct a transcriptional regulatory sub-network involving five SCW-associated *Eucalyptus* TFs. In our approach, we considered well-described candidates such as EgrMYB1, a SCW repressor (closest homolog of AtMYB4), EgrMYB2, a SCW master regulator (functional ortholog of AtMYB83), and EgrMYB137, a TF associated with SCW thickness and vessel density,

as well as lesser-known regulators such as EgrMYB135 (the closest homolog of AtMYB52) and EgrMYB122 (closest homolog of AtMYB43). Additionally, we interrogated some of the technical factors that possibly affect DAP-seq data quality, identified genome-wide binding sites of *Eucalyptus* R2R3-MYB TFs linked to SCW regulation, and developed a machine learning (ML) classifier trained on *Arabidopsis* DAP-seq, TF perturbation, open chromatin and conserved noncoding sequence data for improved target gene identification and physical network inference using DAP-seq.

Materials and methods

DNA construct preparation

Coding sequences of *EgrMYB1* (Eucgr.G01774), *EgrMYB2* (Eugr.G03385), *EgrMYB122* (Eucgr.J01601), *EgrMYB135* (Eucgr.K02297) and *EgrMYB137* (Eucgr.K02806) were chemically synthesised (US Department of Energy Joint Genome Institute), cloned into pCR8©/GW/TOPO™ (Invitrogen, Carlsbad, CA, USA) and then transferred using Gateway™ LR Clonase™ II Enzyme mix (Invitrogen) to the pIX-HALO expression vector (ARABIDOPSIS INTERACTOME MAPPING CONSORTIUM 2011).

Protein expression and detection

HALO fusion proteins were expressed using the TNT SP6 Coupled Wheat Germ Extract System (Cat. # L5540) or the TnT® Coupled Reticulocyte Lysate System (Cat. # L4600) (Promega, Madison, USA) following the manufacturer's specifications for

a 50 µL reaction containing ~800 ng plasmid DNA with a 1-2 hr incubation at 30°C. We then carried out SDS-PAGE followed by a qualitative western blot to validate the corresponding band sizes of the proteins. A commercial anti-HALOTag® rabbit polyclonal antibody (Cat. # G9281) was used as a primary antibody (Promega, Madison USA) together with a Pierce® Goat anti-Rabbit IgG (H+L) coupled to Peroxidase Conjugate (Cat. #G-21234) (Thermo Fisher) for chemiluminescent detection. For optional quality control (QC), performed only for EgrMYB137, HALO fused proteins were detected by western blot using 10 µl of crude Promega TNT® Quick Coupled Transcription/Translation System, 10 µl of the supernatant after protein-DNA interaction step and 5 µl of beads coated with EgrMYB137-HALO proteins.

Genomic DNA library preparation & DNA binding assays

Developing secondary xylem of field-grown *E. grandis* clone TAG0014 (Mondi Tree Improvement Research, South Africa) was sampled in KwaMbonambi, South Africa, immediately flash-frozen in the field and stored at -80°C. Independent genomic DNA samples were extracted using the NucleoSpin Plant Kit (Machery-Nagel) and used to prepare minimal adapter-ligated genomic DNA templates for DAP-seq analysis at the University of Pretoria, named E1 and E2, and University of Toulouse, named E3 and E4. Genomic DNA library preparation, DNA affinity purification, library amplification (introduction of indexes) and library pooling were all carried out as previously described (Bartlett et al. 2017) with minor modifications (see Supplementary Methods S1). Two PCR cycle points from the same binding reaction (15 & 20 cycles) were used following DAP-seq binding assays of each of the E1 and E2

adapter-ligated template samples (Supplementary Table S1). Sequencing was carried out on either an Illumina NovaSeq 6000 (PE 150) platform (Novogene Inc., Sacramento, USA) or an Illumina HiSeq 4000 (PE150) (GeT-PlaGe, Toulouse, France). Optional QC steps were performed only on E3/E4 libraries (Toulouse) to check post-binding qPCR validation and pooled libraries size selection, as described in Figure 1. DNA profiles were analysed by gel and capillary electrophoresis (using a Fragment Analyzer System®, Agilent). SPRIselect bead-based Double Size Selection (Beckman Coulter, Brea, CA, USA) was performed on pooled DNA libraries as per manufacturer's instructions using a ratio of 0.85x-0.56x. Before size selection, we eliminated libraries with irregular DNA profiles and/or low amplification rate to achieve an equimolar pooling of 3 independent DAP experiments per TF (see Supplementary Methods S1).

Bioinformatics and statistical analysis

A FASTQC (Andrews 2010) analysis was carried out followed by read trimming using BBDuk (https://jgi.doe.gov/). Clean reads without adapter contamination were mapped to the *E. grandis* genome v2.0 (Myburg et al. 2014) using Bowtie2 software (Langmead et al. 2009) which was followed by the removal of duplicate reads. Using the mapped reads from unfused HaloTag as controls (to remove background noise), binding regions (peaks) were called using MACS2 and GEM (Zhang et al. 2008; Guo et al. 2012). Peak reproducibility was assessed using the Irreproducible Discovery Rate (IDR) method, while the signal-to-noise ratio was assessed using Fraction of Reads in Peaks (FRiP) values (Li et al. 2011; Landt et al. 2012). Enriched motif discovery was done using MEME (Bailey and Elkan 1994; Bailey et al. 2009). The

overrepresentation of Gene Ontology categories was assessed using BiNGO, a Cytoscape plugin (Maere et al. 2005).

Machine learning classifier for target gene identification

Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) classifiers were first trained and validated on 14 published *Arabidopsis* TF datasets, for which both DAP-seq binding sites (O'Malley et al. 2016) and TF direct genes targets inferred from transient overexpression in the presence of cycloheximide (O'Malley et al. 2016; Brooks et al. 2019) were available, as described in Supplementary Methods S2. After assigning potential target genes using proximity-based methods (ChIPpeakAnno; (Zhu et al. 2010)), true target genes were labelled as such if they showed evidence of differential expression following TF perturbation (Brooks et al. 2019). Three equal negative training sets were constructed: a random selection of remaining false target genes (RANDOM), those with undetected expression in the profiled tissue (UDG) and those with low expression in the profiled tissue (LEG). A 22-component feature matrix was constructed from *Arabidopsis* DAP-seq (O'Malley et al. 2016), DNase-seq (Sullivan et al. 2015), conserved noncoding sequence (CNS) (van de Velde et al. 2016) and co-expression data across a diverse set of transcriptomes for each TF-gene pair (Supplementary Table S2). ML algorithms were evaluated with respect to classification algorithm, choice of negative training set, feature importance and various groups of input features by evaluating the area under the receiver operating characteristic curve (AU-ROC), using five-fold cross-validation and independent training and validation datasets (Supplementary Methods S2). The optimal classifier was directly transferred to *Eucalyptus*, whereby

a similar feature matrix was constructed from the DAP-seq data in this study and published chromatin, co-expression and noncoding sequence data (Supplementary Methods S2).

EgrMYB137 functional characterization

The *EgrMYB137* coding sequence was amplified from *E. grandis* xylem cDNA using Phusion high fidelity Taq polymerase and gene-specific primers listed in Supplementary Table S3. Using the Gateway system, it was first cloned into the entry vectors pDONR207 and pENTR/D-TOPO through BP reaction and then transferred to the different destination vectors through LR reaction, as previously described by Soler et al. (2016). The destination vector pFAST-G02 (Shimada et al. 2010) was used for overexpression in *A. thaliana*; pH35GEAR vector (courtesy of Taku Demura, NAIST, Nara, Japan) for overexpression as dominant repression chimeric proteins in *Arabidopsis* and *Populus*; and pGWAY-0 vector (Plasencia et al. 2016) for overexpression in *Eucalyptus* hairy roots. pFAST-G02-*EgrMYB137* and pH35GEAR-*EgrMYB137* were introduced into *Agrobacterium tumefaciens* strain GV3101/pMP90 (Koncz and Schell 1986) through heat shock as described by Soler et al. (2016) and pGWAY-0-*EgrMYB137* was introduced into *Agrobacterium rhizogenes* strain A4RS as previously described (Plasencia et al. 2016). *Arabidopsis* col-0 plants were transformed using pFAST-G02-*EgrMYB137* and pH35GEAR-*EgrMYB137* constructs as previously described (Soler et al. 2016) and transgenic lines were selected based on transgene expression (Supplementary Figure S1). Plants were grown under short-day conditions (9h light-15h dark) at 25°C in a growth chamber for eight weeks. The base of the inflorescence stem was collected in 80% ethanol for histochemical

analysis while the rest of the inflorescence without leaves and siliques was immediately frozen in liquid nitrogen, then milled to powder using a ball-mill (MM400, Retsch) and kept at -80°C.

Hybrid poplar *Populus tremula x P. alba* (INRA clone 717-1-B4) was transformed with the pH35GEAR-*EgrMYB137* construct and grown in greenhouse conditions as described by Soler et al. (2016). A 5 cm portion at the base of the stem was sampled from 3-month-old poplar plants, and kept in 80% ethanol for histological analysis while the rest of the stem was debarked, frozen in liquid nitrogen, milled to powder and kept at -80°C.

Transgenic *Eucalyptus* roots overexpressing *EgrMYB137* were generated through *Agrobacterium rhizogenes* mediated transformation of *E. grandis* seedlings, as described in Plasencia et al. (2016). Plants were subsequently grown in 200 ml pots for 5 months (16h/12h, 25/22°C) in OIL DRI substrate US-Special Substrate (Type III/R, Damolin, Fur, Denmark). From transformed roots, selected based on DsRed fluorescence, 5 mm long fragments of the main root situated immediately below the crown were collected and kept in 80% until use while the rest of the root was frozen in liquid nitrogen, milled to powder and kept at -80°C.

To investigate the transcriptional activity of the EgrMYB137 protein, the full-length coding sequence of EgrMYB137 was cloned into a pGBKT7-GW vector (Addgene, Inc) containing the GAL4 DNA binding domain. The negative control pGBKT7 and pGBKT7-EgrMYB137 were transferred into yeast strain AH109 using the PEG/lithium acetate transformation procedure. The transformants were cultured in

a synthetic defined medium with SD/-Trp and incubated at 30°C for 3 days. Positive clones were first confirmed by PCR and then grew on SD/-Trp plates. Positive clones of both negative control and pGBKT7-EgrMYB137 were picked for propagation in a 2 mL SD/-Trp liquid medium. Transcriptional activation was determined by transferring them to a synthetic defined medium (SD/-Trp/-His/) containing X-α-gal with incubation at 30°C for 3 to 5 days.

For histochemical staining, transverse sections of stems (90 µm, *Arabidopsis* and poplar) and roots (40 µm, *Eucalyptus*) were obtained, stained, and observed as described in Soler et al. (2016). Leaves from Pro35S:EgrMYB137-EAR and empty vector control poplar lines were stained with Phloroglucinol-HCl and immediately observed with a stereomicroscope Axiozoom V16 (Zeiss, Germany). Lignin content was evaluated on extractive-free samples obtained as described by Soler et al. (2016), using the acetyl bromide lignin method (Pitre et al. 2007). Gene expression analyses in pH35GEAR-*EgrMYB137* transgenic poplars was performed as described in Soler et al. (2016). Briefly, after RNA extraction from poplar debarked stems, we performed DNAse treatment and reverse transcription to obtain cDNA samples. We used the Biomark® 96.96 Dynamic Array platform (Fluidigm) to assess SCW-related gene expression by microfluidic RT-qPCR as described in Cassan-Wang et al. (2012). Transcript abundance was calculated with the $2^{-\Delta\Delta Ct}$ method, with the geometric mean of five validated housekeeping genes to normalize the results: Ubiquitin (UBQ), CDC2, HK1, HK3, and HK11 (Legay et al. 2010; Sixto et al. 2016). Primers sequences are listed in Supplementary Table S3.

SCW chemical fingerprinting was performed by FT-IR coupled to multivariate statistical analyses (see Supplementary Methods S3) using xylem samples harvested from *Populus* and *Eucalyptus* lines beforehand washed to eliminate soluble extractives as described in Ployet et al. (2019).

Results

Optimisation of the DAP-seq workflow and evaluation of transcription factor binding site reproducibility

The protocol of DAP-seq, applied to *Eucalyptus* MYB TFs in the present work, was adapted from the method described in Bartlett et al. (2017). A schematic view of the key steps is represented in Figure 1. Briefly, a sonicated genomic DNA library (gDNA) was generated from *E. grandis* developing xylem tissues, ligated to sequencing adaptors, and purified using magnetic beads coated with HALO-tagged MYB TF translated *in vitro*. Purified gDNA fragments bound on MYB TFs were washed, isolated, released, amplified, and sequenced to identify TFBSs. In order to reduce potential technical limitations that could explain the many unsuccessful DAP-seq assays reported in previous studies using large-scale screens (O'Malley et al. 2016; Bartlett et al. 2017), we added additional QC to improve DAP-seq efficiency and reproducibility (Figure 1a), as well as providing qualified improved standards for independent DAP-seq experiment comparison.

Optional QCs concern (1) the detection of HALO-TF protein successfully bound on beads and (2) the measurements of gDNA library enrichment at critical steps of the

DAP-seq protocol (Figures 1b to e). First, in addition to the input quantity of HALO-TF fusion protein, the quantity of TF bound to beads through HALO-ligand (chloroalkane) recognition likely influences DAP-seq success. As shown in Figure 1C (left panel), we used an anti-HALO antibody to detect EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135, and EgrMYB137 fused to the HALO-tag in the *in vitro* translation reaction mix, along with a HALOTag positive control. We demonstrated that a fraction of TF-bound beads can be used to assay TF binding on beads (Figure 1C, right panel) as an additional quality step during the DAP-seq assay and is appropriate in case of unsuccessful protein detection after *in vitro* production or unbound HALO-protein signal in the supernatant. Second, in addition to PCR quality controls proposed by Bartlett et al. (2017), efficiency and homogeneity of DNA recovery was assessed by capillary electrophoresis and/or qPCR before and after the steps of DNA sonication, adaptor ligation, DNA amplification, and library purification. A fast and efficient beads-based DNA purification method (SPRIselect®) was compared with gel purification. SPRIselect increased DNA recovery yield after library pooling and lowered the amount of 100-150 bp dimer contaminants detrimental for next-generation sequencing.

The reproducibility of DAP-seq data has not been well studied, and best practices for identifying reproducible TFBSs from DAP-seq data have not yet been established. Therefore, we implemented DAP-seq analysis of EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135, and EgrMYB137 across two independent laboratories, each with technical replication, and additionally with inter-lab replication for EgrMYB1 and EgrMYB2. In Laboratory 1, technical replication involved two independently

prepared pre-binding gDNA libraries prepared as per Bartlett et al. 2017, with additional evaluation of different PCR cycle numbers during full-length adapter indexing to evaluate the effect of this parameter on sequence yield and duplication levels. In total, ~173 million paired clean reads were obtained from laboratory 1 (Supplementary Table S1). After testing different PCR cycle numbers during library preparation, we observed that 15 PCR cycles reduced the proportion of duplicated sequence reads but overall yielded lower quantities of total reads than 20 PCR cycles (Supplementary Table S1). Therefore, we opted to continue with 20 cycles. Additionally, substantial differences were observed between DAP-seq experiments performed on independent pre-binding gDNA libraries within the same experiment, where one library (E2) yielded significantly more sequence data than the other (E1) (Supplementary Table S1), despite yielding similar qPCR values in pre-binding library quality control as advised by Bartlett et al. (2017) (data not shown). This suggested that post-binding quality control (e.g., qPCR validation) could ensure low-quality libraries are discarded prior to sequencing.

In Laboratory 2, EgrMYB1, EgrMYB2 and EgrMYB137 DAP-seq datasets each included three technical replicates (Supplementary Table S1). To reduce library failure, we included additional capillary electrophoresis to monitor gDNA fragments of interest along DAP steps, and additional qPCR analyses (Figure 1b, d). As advised by Bartlett et al. (2017), the average size of gDNA fragments was around 200 bp after sonication, 260 bp after minimal adaptor ligation (Figure 1b), and 320 bp after PCR amplification using full-length adaptors (Figure 1d). DNA purification dramatically decreased gDNA template (Figure 1d): Ct rose from 7.3±0.8 for ligated gDNA to

24.4±2.7 for recovered DNA. The highest variability of gDNA template among and between library replicates was also observed at this step, but library homogeneity was restored post-PCR (Figure 1d). Ligated-DNA amplification by PCR generated a high amount of detrimental 80-150 bp DNA fragments, likely synthesised from free adaptors (black arrows, Figure 1e). Because Illumina technology is biased towards the sequencing of small fragments, we tested a bead-based post-DAP gDNA size selection, which enabled an efficient elimination of contaminating dimers while reducing gDNA loss compared to classical gel purification (Figure 1e).

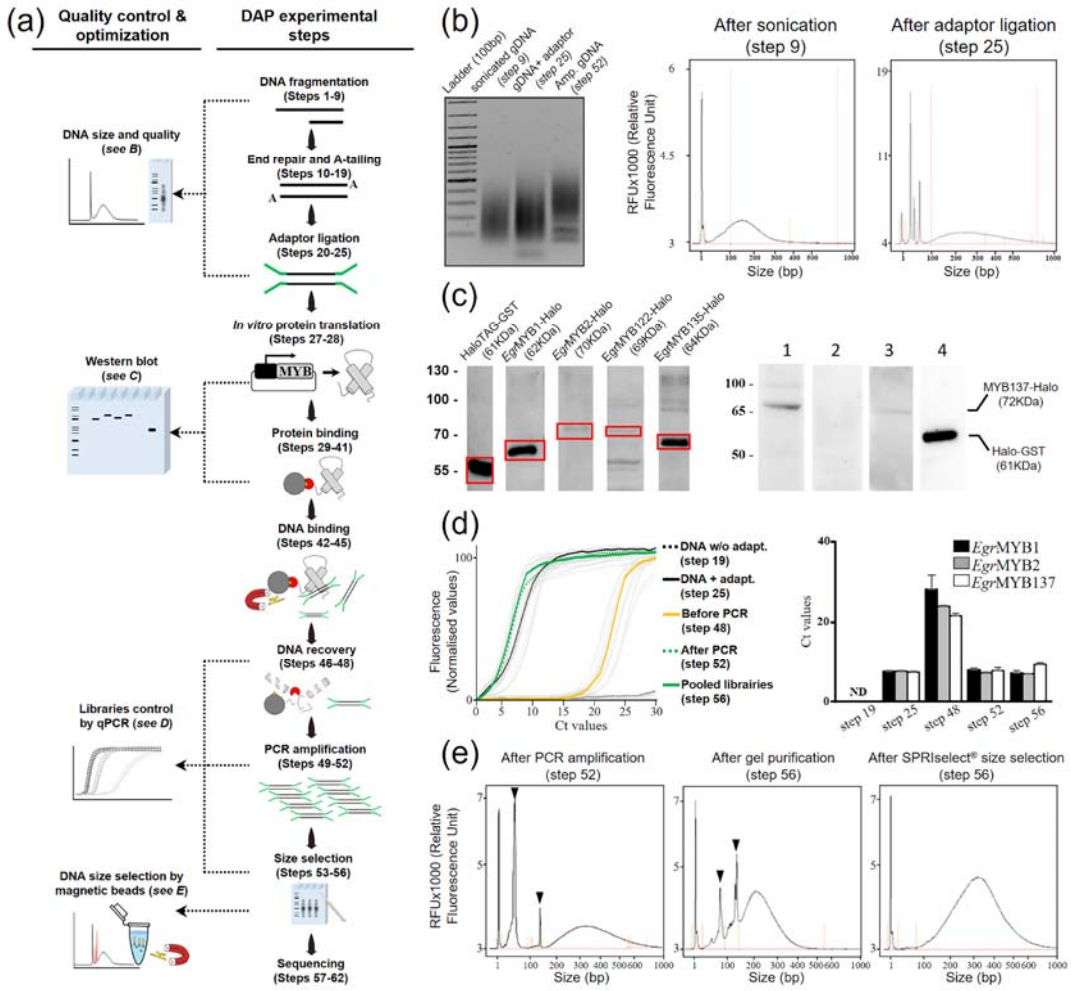


Figure 1: DAP-seq protocol optimization and quality controls. (a), Overview of DAP-seq protocol. Steps are numbered as described in Bartlett et al. (2017). Quality control and optional optimization steps are presented on the left. **(b)**, DNA profiles were analysed by gel and capillary electrophoresis. Average sizes of approximately 200 bp and 260 bp are expected after fragmentation (step 9) and minimal adaptor ligation (step 25), respectively. **(c)**, Detection of EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135 and EgrMYB137 by Western blot after *in vitro* protein translation (left panel). An optional step of Western blot, using EgrMYB137 as an example (right panel), was performed to confirm MYB-HALO proteins binding on magnetic beads. The correct production of EgrMYB137-HALO fused proteins was assessed at step 28 after *in vitro* translation (lane 1). The fixation of EgrMYB137-HALO on beads is confirmed by comparing, at step 48, EgrMYB137-HALO weak/missing signal in the supernatant (lane 2) and a strong signal using protein-bound beads (lane 3). HALO-GST tag was used as a positive control (lane 4). **(d)**, Efficiency and homogeneity of adaptors-ligated gDNA template recovery was assessed by qPCR at steps 19, 25, 48, 52 and 56. Examples of amplification curves (left panel) and Ct values (right panel) are presented for EgrMYB1, EgrMYB2 and EgrMYB137 libraries. Bold curves in the legend are the mean of technical replicates (background shaded curves). **(e)**, DNA size selection was compared between gel purification and SPRIselect magnetic beads selection. PCR-based amplification of purified gDNA produces a high quantity of undesirable small 80-150bp fragments (black arrows). DNA profile analyses demonstrated that bead-based selection was more efficient to discard small fragments (middle and bottom panels) that are detrimental to Illumina sequencing, and ensured a higher yield of selection of 320 bp DNA fragments (expected size after full adaptors ligation, step 52).

DAP-seq analysis was replicated within and, for EgrMYB2, between independent laboratories. We applied three quality metrics to assess libraries: first, libraries had to yield at least 1,000 DAP-seq peaks to be considered informative; second, they required a Fraction of Reads in Peaks (FRiP) score of ≥5% as per ENCODE signal-to-

noise recommendations for ChIP-seq (Landt et al. 2012); third, their peaks had to exhibit enrichment for the expected AC-I (ACCTACC), AC-II (ACCAACC), or AC-III (ACCTAAC) canonical motif (Hatton et al. 1995; Raes et al. 2003; Patzlaff et al. 2003; Zhong et al. 2008; Zhou et al. 2009; Zhong and Ye 2012). We called peaks using MACS2 (Zhang et al. 2008) and GEM (Guo et al. 2012) to compare the two peak callers. For libraries meeting the criteria, GEM generated more peaks than MACS2 (an average of 11,421 peaks per candidate for GEM versus 3,819 peaks per candidate from MACS2) and had a higher average FRiP score (21.6% versus 19.5%) (Supplementary Table S4). Therefore, we proceeded with GEM for peak calling. Based on our stringent criteria, five of the seven replicates for EgrMYB2 but only seven of sixteen libraries from EgrMYB1, EgrMYB122, EgrMYB135 and EgrMYB137 passed these criteria (Supplementary Table S4 and Supplementary Table S5), corresponding to a per-library success rate of 52%. By similar metrics, ours is a 3-fold improvement over the 17.3% (314 of 1,812 TFs) subjected to DAP-seq by O'Malley et al. (2016) that were successful. Furthermore, if we consider the additional QC steps implemented in Laboratory 2, our per-library success rate was ~78%.

For EgrMYB2, two technical replicates from Laboratory 1 (EgrMYB2_E1_15 and EgrMYB2_E1_20) had fewer than 1000 peaks each and FRiP scores below 5% and hence discarded (Supplementary Table S6). Since their peaks were still enriched for AC-II elements, the affinity enrichment of these libraries was likely successful, but the sequence coverage was insufficient to yield high peak numbers. The number of peaks in common between EgrMYB2 replicates ranged from 3,823 (Laboratory 2) to 15,263 (Laboratory 1) (Supplementary Table S6). To determine the final peak set for

EgrMYB2 we merged the peak sets from each laboratory and took the overlapping set of peaks between laboratories. In doing so, only high confidence peaks were retained. The best pairwise overlap of peaks between replicates for EgrMYB135 and all three replicates for EgrMYB137 that passed our criteria were similarly merged. Finally, only single replicates for EgrMYB1 and EgrMYB122 yielded over 1,000 peaks and FRiP ≥5%. Therefore, while DAP-seq success rates are still far from 100%, by including multiple technical replicates per candidate we were able to discard poor replicates and obtain high-quality data for all five of our candidates.

Development of a random forest (RF) classifier for target gene assignment

In previous work, we observed that putative target genes linked to SCW-related EgrMYB DAP-seq binding sites located in open chromatin were overrepresented in systems biology datasets linked to wood formation biology, while those found outside of accessible chromatin were not (Brown et al. 2019). This suggests that true target genes could be distinguished from incidental DAP-seq peak associations based on complementary data that provide functional context. We sought to build on this premise and improve the accuracy of proximity-based target gene assignment by leveraging high-confidence target gene-DAP-seq peak associations and various functional omics data in *Arabidopsis* (Supplementary Table S2) to train and transfer a supervised ML classifier to similar features calculated from *Eucalyptus* data.

Brooks et al. (2019) identified direct target genes of 33 nitrogen-early response TFs in *Arabidopsis* root cells from protoplast-based post-translational induction assays

(Bargmann et al. 2013). We used this data to train a random forest (RF) classifier to discriminate *bona fide* gene targets proximal to DAP-seq binding events by labeling candidate TFBS-target pairs as true (i.e., transient overexpression of the TF causes differential expression of the target gene) or false (no differential expression following transient TF induction). Of 33 TFs analysed by Brooks et al. (2019), 14 had DAP-seq data available (O'Malley et al. 2016) for classifier training. Classifier performance was evaluated for different classifiers trained on three negative sample sets comprising genes with undetected (UDG) or low (LEG) expression values versus random selection of non-differentially expressed genes. The rationale for this was that genes with undetected or low expression in the same tissue as a TF of interest are least likely to be true targets and include fewer false negatives (Song et al. 2020). One-way ANOVA analyses showed that the choice of algorithm and the negative training dataset significantly affected classifier performance, where post-hoc testing for AUC-ROC scores revealed that the RF classifier developed on matrices with UDGs as negative samples performed better than the other methods (Figure 2a; *P*-value < 0.05).

An absolute pairwise correlation threshold of 0.7 was used to identify redundant features (Supplementary Figure S2). After removing the least important feature of each pair based on RF feature importance values, eleven features remained. Of these, the Pearson correlation coefficient (PCC) showed the highest RF feature importance, while features relating to DNase-seq and CNS data were the least important (Supplementary Figure S2). This suggests that the co-expression relationship between each TF and a possible target gene contributed the most information to the classifier. Removing the eleven redundant features reduced the performance of the

RF classifier by a small degree (decrease in AUC-ROC of 0.018 and 0.019 for UDGs and LEGs negative sample sets, respectively). To better understand what categories of features were important for classifier performance, the RF classifier was evaluated on training and testing matrices with different combinations of features. The various feature sets all had a statistically significant effect on the AUC-ROC performance of the classifier (one-way ANOVA, *P*-value < 0.05) (Figure 2b). RF classifier performance using seven features related to DAP-seq data exclusively was favourable compared to using only the co-expression data feature (*P*-value < 0.05, Tukey HSD), while integrating DAP-seq and co-expression learning features improved classifier performance when compared to using either of these feature sets alone (*P*-value < 0.05, Tukey HSD). Including features relating to DAP-seq peak overlap with open chromatin and/or CNS regions resulted in only a modest improvement in classifier performance with or without expression data already included as features (Figure 2b). With this result, we proceeded with the set of eleven non-redundant features.

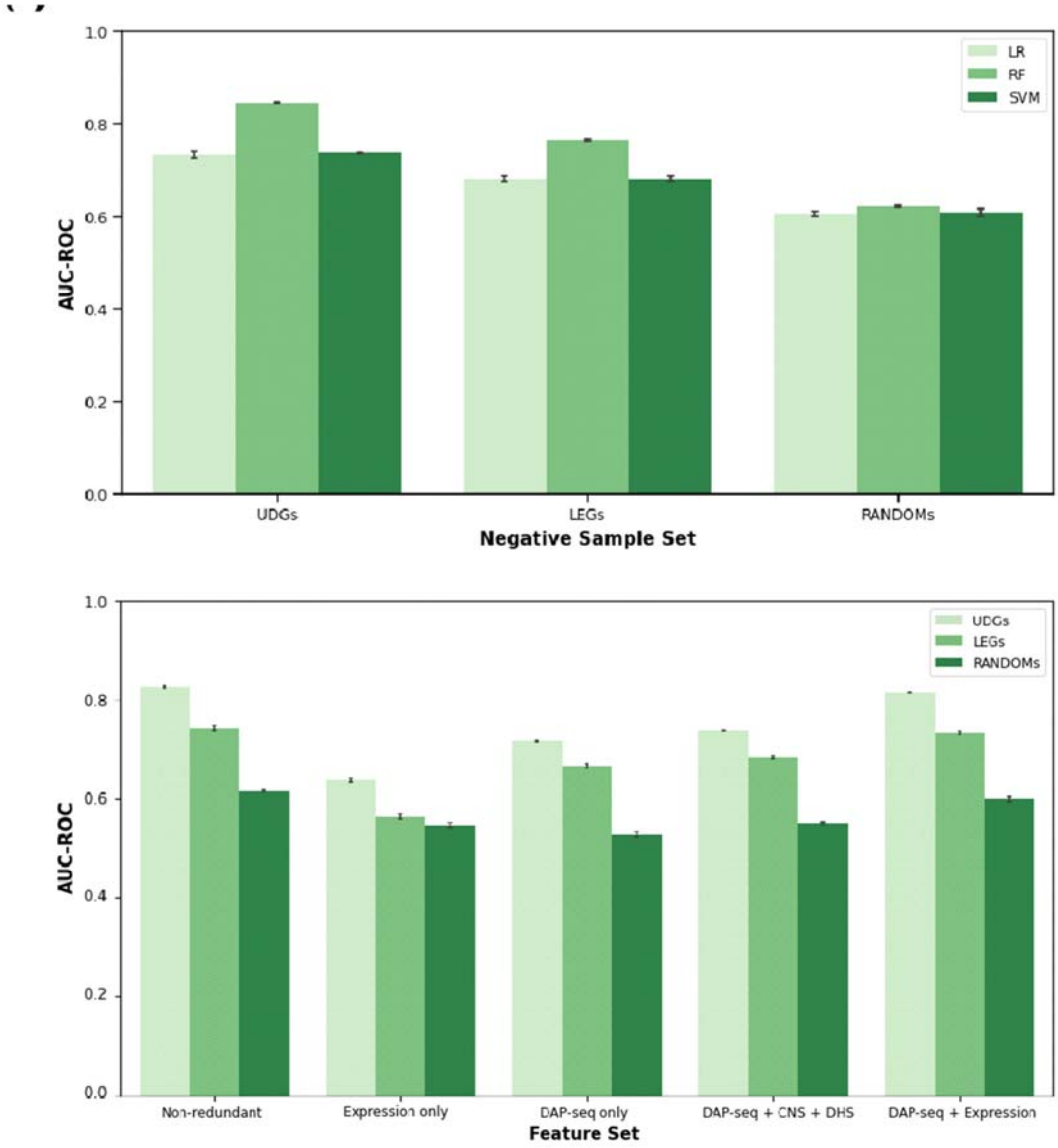


Figure 2. Performance of DAP-seq peak-target gene classifiers in *Arabidopsis thaliana*. (**a**) Evaluation of logistic regression (LR), random forest (RF) and support vector machine (SVM) classifiers. (**b**) Performance of RF classifiers trained on different clusters of features. UDGs, non-differentially expressed genes with undetected expression in roots; LEGs, lowly expressed genes; RANDOMs, randomly selected non-differentially expressed genes; CHS, conserved noncoding sequences; DHS, DNase I hypersensitive sites. Lowercase letters on each bar represent AUC-ROC scores that are significantly different from each other based on post hoc testing (Tukey HSD test, p-

value <0.05). Error bars represent one standard deviation and show variability in the AUC-ROC across the best-performing models for each of five cross-validations.

Table 1. Evaluation of conventional distance-based methods and the random forest classifier trained on different negative samples

	Random forest classifier			Distance-based method	
	UDGs[a]	**LEGs[b]**	**RANDOM[c]**	**Nearest gene to each DAP-seq peak**	**DAP-seq peak within 5 kb of gene TSS[d]**
Precision	0.751	0.689	0.581	0.105	0.098
Recall	0.711	0.723	0.595	0.243	0.377
F1 Score	0.730	0.705	0.588	0.147	0.156

[a]UDGs: classifier trained on genes with undetected expression for the target tissue as negative samples.

[b]LEGs: classifier trained on lowly expressed genes (TPM < 5) for the target tissue as negative samples.

[c]RANDOMs: classifier trained on randomly selected non-differentially expressed genes as negative samples.

[d]TSS, Transcription Start Site

In evaluating the performance of the RF classifier against conventional distance-based target gene assignment, two baseline methods were considered. The RF classifier showed substantially higher precision, recall, and F1 scores than both the proximity-based approach that assigned the nearest gene to a DAP-seq peak, as well as one that assigned genes to DAP-seq peaks if the TSS was within 5 kb of the peak (Table 1; Supplementary Table S7). This was the case for all three types of negative training sets (UDGs, LEGs, and random selection). The RF classifier was between 5.5 and 7.7-fold more precise than distance-based methods and 1.6 to 3-fold more accurate than either baseline, particularly for the RF classifier trained on datasets

with UDGs as negative samples which exhibited an F1 score 4.7 to 5-fold higher than the distance-based methods (Table 1). Overall, the RF classifier predicted a much greater proportion of true TF-target associations with superior accuracy.

To assess the classifier's robustness when applied to independent datasets, the RF classifier was iteratively re-trained and tested on independent training and testing submatrices selected from the 14 *Arabidopsis* TF datasets from Brooks et al. (2019). Five independent training/testing data splits were considered such that all datasets from the 14 TFs were used while maintaining an approximate 80/20 training/testing split, a balanced number of positive and negative samples in the training set and ensuring that samples from different TF families were in either the training or testing subsets but never both (Supplementary Table S8). One-way ANOVA analysis showed that the way in which the data was split had a negligible but significant effect on the performance of each of the negative sample sets (Supplementary Figure S3, *P*-value < 0.05). Despite the modest decrease in AUC-ROC metrics, the data indicate that the RF classifier performed robustly on independent DAP-seq datasets, with an AUC-ROC of ~0.72 when considering UDGs as the negative sample set. Since the latter was consistently superior compared to other negative sample sets, we retained the RF classifier trained on UDGs for gene target prediction from DAP-seq data.

We next implemented the RF classifier in *Eucalyptus* to improve gene target assignment of the final peak set of EgrMYB2 by direct transfer of the classifier using a similar set of 11 non-redundant *Eucalyptus* features. We obtained a total of 9,626 possible target genes using proximity-based peak assignment. Using the ML classifier

and applying a probability (Pr) threshold Pr ≥ 0.5, we retained only 2,108 target genes and a mere 906 target genes at P ≥ 0.7. Using Gene Ontology (GO) analysis of the assigned target genes to gain insight into the possible biological roles that EgrMYB2 plays in woody plants such as *Eucalyptus*, we looked at the fold change of overrepresented GO terms and assessed the enrichment of some of these terms at the three cut-off points in comparison to the baseline method (pre-ML) (no cut-off, Pr ≥ 0.5 and Pr ≥ 0.7) as shown in Supplementary Table S9. Prior to the ML step, there was a significant overrepresentation of GO terms related to phenylpropanoid and lignin biosynthesis, in broad agreement with altered lignin profiles observed in EgrMYB2-overexpressing tobacco (Goicoechea et al. 2005). To examine the effect of the ML algorithm, we carried out GO enrichment analysis for post-ML datasets and observed even stronger enrichment for phenylpropanoid and lignin-related terms, but also a large number of (secondary) cell wall GO terms among the top-ranking terms pre-filtering (Supplementary Table S9), consistent with the ability of EgrMYB2 to increase SCW thickness (Goicoechea et al. 2005) as well as the SCW master regulator status of its homologs AtMYB46/83. This suggests that the ML algorithm is filtering incidental and false positive targets but retaining biologically plausible DAP-seq target gene associations.

Identification of EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135, and EgrMYB137 target genes

Despite more than a decade of research on the biological roles of EguMYB1 and EguMYB2 from *Eucalyptus* (Goicoechea et al. 2005; Legay et al. 2007; Legay et al.

2010), most of their target genes remain unknown. We set out to identify genome-wide binding sites for these candidates in *E. grandis* along with EgrMYB122, EgrMYB135, and EgrMYB137, as well as their target genes. We observed that the AC-rich element known for MYB-family TFs (O'Malley et al. 2016) was prominent in the peaks sets of all five TFs and highly similar to their closest *Arabidopsis* homologs (Supplementary Table S5). This result suggests that the DNA-binding specificity of these putative orthologs are conserved.

Using our optimised target gene ML classifier trained on *Arabidopsis* data and transferred to *Eucalyptus*, we inferred gene targets for all five SCW-related TF candidates. When we analysed the binding profiles of each TF relative to the transcription start site (TSS), we observed that EgrMYB2, EgrMYB135 and EgrMYB137 DAP-seq peak frequencies increased dramatically within 1,000 bp immediately upstream of the TSS, dropping again near the TSS and then increasing again over much of the gene body (Figure 3a). Interestingly, EgrMYB1 and EgrMYB122 binding profiles, which were similar to each other, showed a preference for binding further upstream of the TSS (-4 kb to -1 kb), with infrequent binding occurring over most of the gene body (Figure 3a). These differences in *cis*-regulatory architecture underpin the importance of incorporating such features into the target gene assignment process. We obtained an average of 7,059 target genes per candidate pre-ML using the baseline method of assigning peaks to proximal genes. We then used the confidence probability score from the classifier to filter possible false positive associations. Post-ML (but without a probability cut-off), we were left with

an average of 5,894 target genes, further reducing to between 479 and 2661 targets when the cut-off of Pr ≥ 0.5 is implemented (Figure 3b).

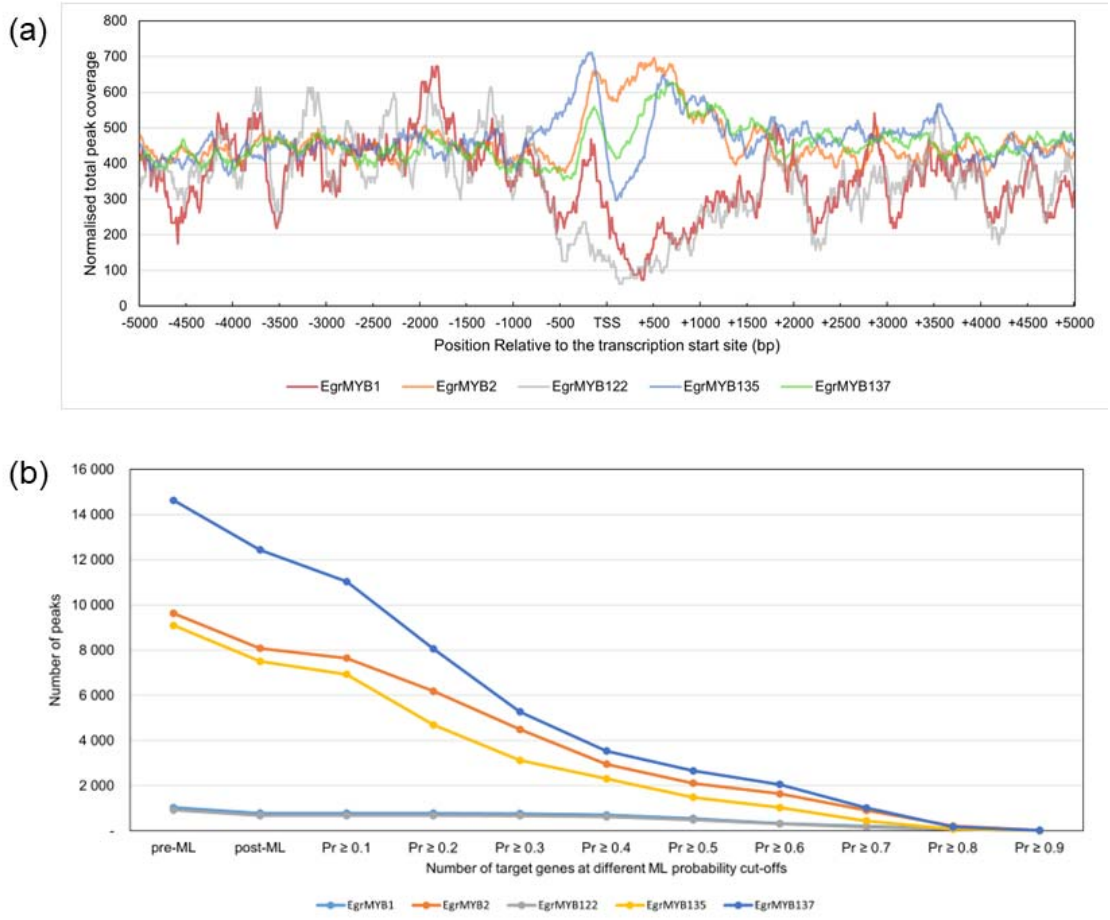


Figure 3. EgrMYB transcription factors binding profiles and their target genes at different DAP-seq-ML probability cut-offs. (a) DAP-seq binding profiles of EgrMYB transcription factors. A typical gene model is indicated for scale. TSS, transcription start site. **(b)** The number of target genes at different probability cut-offs linked to the ML classifier.

To gain insight into the potential biological role being played by each of these candidates at different ML probability thresholds, we carried out Gene Ontology (GO) enrichment analysis on target genes identified with DAP-seq-ML. Biological

processes such as phenylpropanoid metabolism and phenylpropanoid biosynthesis were common and highly enriched across most of the gene target sets even pre-ML, suggesting that these candidates are involved in the regulation of lignin and SCW-related biological process (Supplementary Tables S9-S13). When a ML classifier threshold of Pr ≥0.5 was applied, several additional SCW-related GO terms such as cell wall thickening, SCW biogenesis, lignin biosynthesis, and cell wall polysaccharide biosynthesis were enriched for the transcription repressor EgrMYB1, and activators EgrMYB2 and EgrMYB137, consistent with their known links to SCW regulation (Supplementary Table S9, S10 and S11). This suggests that the ML classifier was successful in removing false positive targets and retaining true targets. Although the enrichment for some of these terms increased further at Pr ≥0.7, a number of terms were lost in the case of EgrMYB1, EgrMYB2 and EgrMYB137, suggesting that an intermediate cut-off was optimal in retaining biologically relevant gene targets (Supplementary Table S9, S10 and S11).

EgrMYB122 target genes, which were not significantly associated with any biological process pre-ML, were strongly enriched in phenylpropanoid pathway-associated GO terms at Pr ≥ 0.5 and Pr ≥ 0.7 (Supplementary Table S12). EgrMYB122 targeted five of the seventeen *bona fide* lignin genes (*EgrPAL9*, *EgrC3H3*, *EgrC4H1*, *EgrCCoAOMT2* and *Egr4CL1*) described by Carocha et al. (2015). Additionally, much of the phenylpropanoid pathway is targeted, as shown by the KEGG pathway analysis (Supplementary Figure S4), suggesting that EgrMYB122 is involved in SCW-related biology. In contrast to the rest of the candidates, EgrMYB135 had no overrepresented GO terms at Pr ≥ 0.5 or Pr ≥ 0.7 (Supplementary Table 13). While several GO terms

were mildly overrepresented in the unfiltered (pre-ML and post-ML) set of genes including phenylpropanoid pathway-associated terms, response to abiotic and biotic stress-related GO terms ranking highest and were the majority of the GO terms represented (Supplementary Table S13). This suggests that although EgrMYB135 might be involved in SCW-related biology, it might also be involved in stress-related biological processes.

Based on what appears to be a sensible DAP-seq-ML threshold of Pr ≥ 0.5, we constructed a SCW-related subnetwork linking the five candidate TFs to *bona fide* cellulose, hemicellulose, as described by Myburg et al. (2014) and lignin structural genes as described by Carocha et al. (2015), as well as SCW-related TFs (see Supplementary Methods S2) (Figure 4). Overall, there were at least 51 TFs, 15 lignin-associated genes, 39 hemicellulose genes, and 5 cellulose-related genes linked to EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135, and/or EgrMYB137 in the network represented in Figure 4. In addition to the core SCW-related cellulose synthases, a substantial number of lignin and hemicellulose genes are targets of these five MYB-family TFs. EgrMYB2 and EgrMYB137 seem to co-target a large number of SCW-related TFs, lignin and hemicellulose genes, suggesting a possible co-regulation of SCW-related biological processes.

Putative orthologs of TFs involved in SCW regulation in *Arabidopsis* were targeted by both EgrMYB2 and EgrMYB137 (Figure 4), including SCW regulators NST1 (EgrNAC49) and SND2 (EgrNAC170), while SND3 (EgrNAC64) and MYB103 (EgrMYB60) are additionally targeted by both EgrMYB2 and EgrMYB137. Additionally, EgrMYB 2 targets SCW regulator SND1 (EgrNAC61; Laubscher et al.

2018) . EgrMYB2, EgrMYB122 and EgrMYB137 target vessel-associated VND1/2 (EgrNAC146). Other TFs were targeted by EgrMYB135 and EgrMYB137, like VND6 (EgrNAC26; (Laubscher et al. 2018)) and REV, EgrMYB2 alone, like AtVND4/5 (EgrNAC50) and EgrMYB122 alone like VND7 (EgrNAC75). EgrMYB2 and EgrMYB137 shared a large number of common SCW-related targets (Figure 4). A total of 62 *bona fide* SCW-related genes and TFs were common between EgrMYB2 and EgrMYB137. At least 15 out of 17 lignin genes involved in 10 steps of lignin biosynthesis previously described Carocha et al. (2015) are shown to be targets in this subnetwork of five MYB-family TFs of which 13 out of the 17 lignin genes are common targets of EgrMYB2 and EgrMYB137.

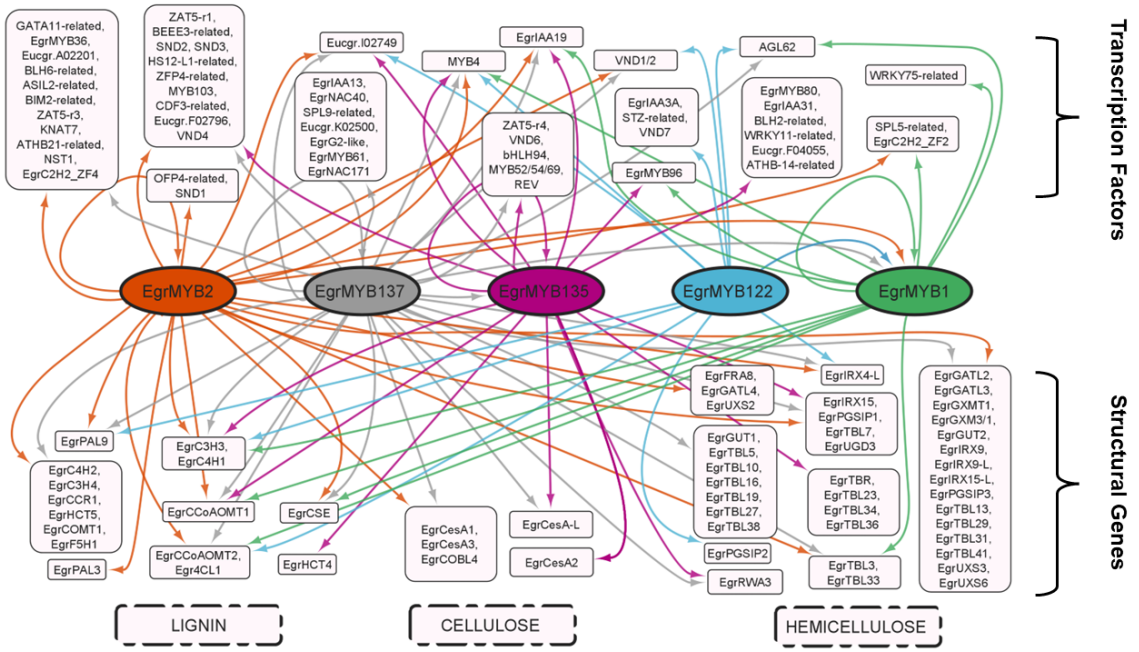


Figure 4. Secondary cell wall-related target gene subnetwork involving EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135 and EgrMYB137. Transcriptional regulatory network of previously described *bona fide* SCW-related structural genes and TFs at (Pr ≥ 0.5), linked (by DAP-seq-ML) to EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135, and EgrMYB137.

These results reveal for the first time the biologically relevant target genes from genome-wide *in vitro* binding sites of the EgrMYB TFs, with biological enrichments observed using GO analysis supporting our hypothesis of a SCW-related role for at least four of the five candidate TFs and EgrMYB135 predicted to target several SCW-related TFs and structural genes. Several overlapping gene targets were observed, with EgrMYB2 and EgrMYB137 sharing many more targets than the other TFs.

Functional characterization of EgrMYB137 in *Arabidopsis*, *Populus*, and *Eucalyptus* transgenic lines

In order to investigate the potential role of EgrMYB137 in SCW formation, we analysed transgenic *Arabidopsis* lines overexpressing either a native form of EgrMYB137 (Pro35S:EgrMYB137) or a dominant-repressive form (Pro35S:EgrMYB137-EAR) fused to an Ethylene-responsive element binding factor-associated Amphiphilic Repression motif (EAR), to transform it into a dominant repressor (Hiratsu et al. 2003). While EgrMYB137 overexpressors display no visible growth phenotype (Figure 5a), the dominant repressors show reduced growth at the rosette stage and a deficiency in stem rigidity (Figure 5a). To clarify the link between EgrMYB137 and wood-related traits, Pro35S:EgrMYB137 overexpressing lines were generated in an *E. grandis* Hairy Roots (HR) system (Plasencia et al. 2016) and Pro35S:EgrMYB137-EAR dominant-repressive lines generated in *Populus tremula x P. alba*. In poplar, Pro35S:EgrMYB137-EAR lines display a floppy stem as previously observed in *Arabidopsis*, and presented curled leaves (Figure 5c). Phloroglucinol-HCl staining on bleached leaves was less intense in vascular tissues in dominant-

repressive lines compared to control (Figure 5d), suggesting a possible reduction of lignin content in leaf vascular tissues of Pro35S:EgrMYB137-EAR poplar lines.

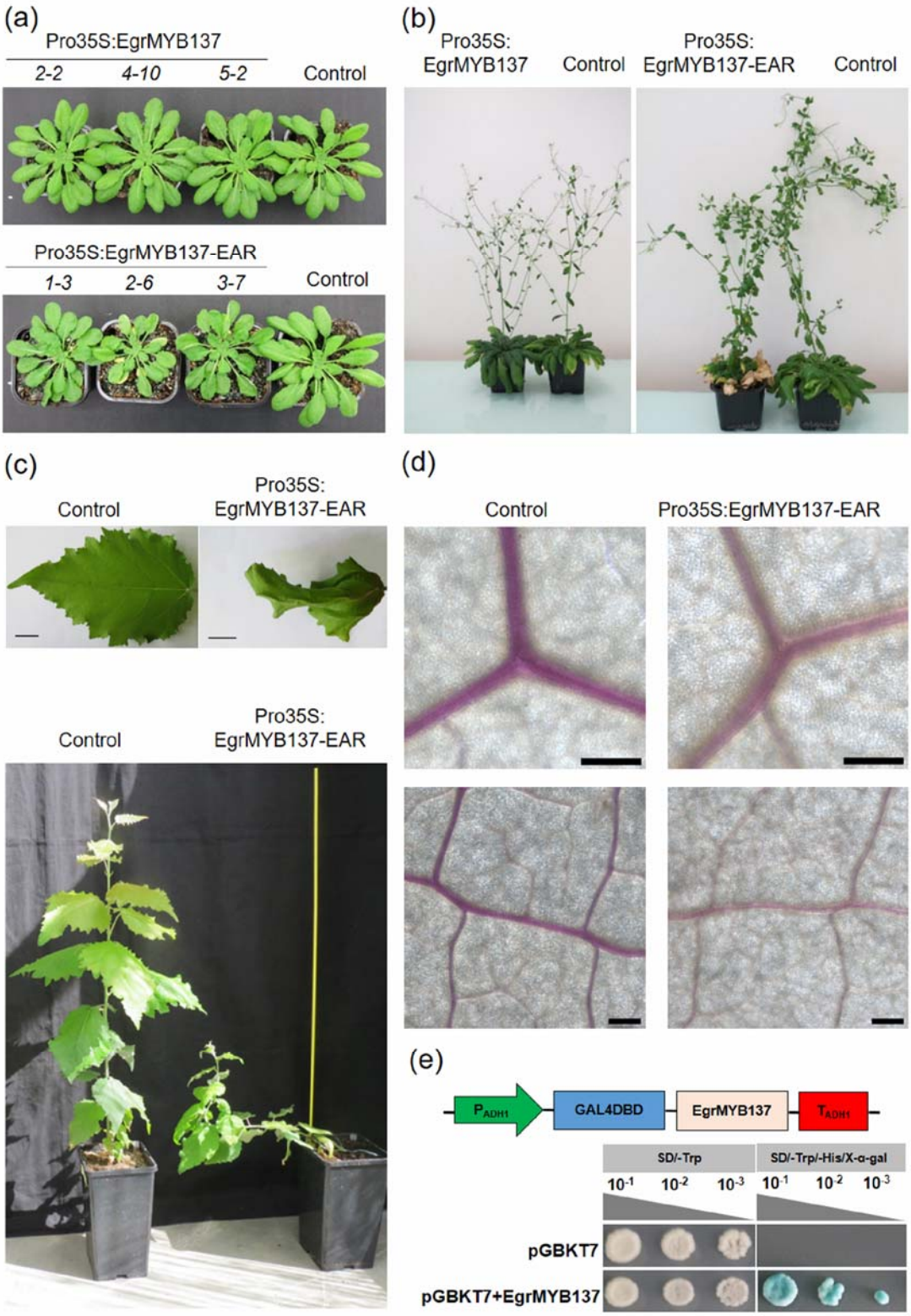

(a)
Pro35S:EgrMYB137
2-2
4-10
5-2
Control
Pro35S:EgrMYB137-EAR
1-3
2-6
3-7
Control
(b)
Pro35S: EgrMYB137
Control
Pro35S: EgrMYB137-EAR
Control
(c)
Control
Pro35S: EgrMYB137-EAR
Control
Pro35S: EgrMYB137-EAR
(d)
Control
Pro35S:EgrMYB137-EAR
(e)
PADH1
GAL4DBD
EgrMYB137
TADH1
SD/-Trp
SD/-Trp/-His/X-α-gal
10-1
10-2
10-3
10-1
10-2
10-3
pGBKT7
pGBKT7+EgrMYB137

Figure 5: Dominant repression of MYB137 alters stem and leaf development in *Arabidopsis* and *Populus*. (a and b) *Arabidopsis thaliana* overexpressing lines EgrMYB137 (Pro35S:EgrMYB137) display no visible growth phenotype when compared to control, while dominant-repressor lines (Pro35S:EgrMYB137-EAR) have a reduced growth and show floppy stems when compared to control. **(c)** *Populus tremula x alba* dominant-repressor lines (Pro35S:EgrMYB137-EAR) display altered leaf morphology and reduced stem rigidity. **(d)** Phloroglucinol-HCl staining of bleached leaves from control and dominant-repressor poplar plants (Pro35S:EgrMYB137-EAR). Pro35S:EgrMYB137-EAR lines show a reduction of phloroglucinol staining in leaf vascular tissue. Scale bar: 0.2 mm. **(e)** Transcription activity analysis of full-length EgrMYB137 in yeast strain AH109. pGBKT7 empty vector served as a negative control. Growth of transformants with pGBKT7 and pGBKT7-EgrMYB137 was assessed on auxotroph medium SD/-Trp (for presence of plasmid) and SD/-Trp/-His/X-gal (for HIS3 auxotrophic plus LacZ reporter activity).

To determine whether it is a transcriptional activator, EgrMYB137 was fused to the GAL4 DNA-binding domain and introduced into engineered yeast strain AH109 containing GAL4-responsive HIS3 and LACZ reporter genes. The yeast transformed with the negative control pGBKT7 and with pGBKT7-EgrMYB137 grew normally on single dropout medium (SD/-Trp) plates (Figure 5e), indicating their successful transformation. The yeast transformed with negative control pGBKT7 empty vector showed no growth on the (SD/-Trp/-His) auxotrophic medium, while that transformed with the pGBKT7-EgrMYB137 vector grew on SD/-Trp/-His auxotrophic medium and showed galactosidase activity in the presence of X-α-gal (Figure 5e). This strongly suggested that the full length of the EgrMYB137 protein positively regulates transcription.

Transverse sections of *Arabidopsis*, *Populus* and *Eucalyptus* HR transgenic lines overexpressing EgrMYB137 and/or EgrMYB137-EAR were analysed and stained with Phloroglucinol-HCl to investigate vascular tissue lignification. While no clear phenotype appeared for *Arabidopsis* EgrMYB137 over-expressing lines, dominant repressor lines in both *Arabidopsis* and *Populus* presented a weaker phloroglucinol staining compared to control lines (Figure 6a, b), indicating a reduction in lignification. For *Arabidopsis*, this reduction in lignification was more visible in interfascicular fibres. In *Eucalyptus grandis* hairy roots, the overexpression of *EgrMYB137* increased phloroglucinol staining of the xylem fiber cells compared to control (Figure 6c). All these results suggest a positive role for EgrMYB137 in xylem cell lignification, confirmed by biochemical analysis of soluble lignin content using the acetyl bromide method (Ployet et al. 2019). In accordance with histological results, dominant repressor lines (Pro35S:EgrMYB137-EAR) in *Arabidopsis* and *Populus* exhibited a significantly lower lignin content compared to control lines (12.0 ± 3.1 vs 15.4 ± 0.6 and 15.7 ± 1.6 vs 19.5 ± 1.0 respectively, expressed in % of dry weight; Table 2) while overexpressing lines (Pro35S:EgrMYB137) in *Eucalyptus* HR had a higher lignin content (18.6 ± 0.4%) relative to control (16.2 ± 0.6%).

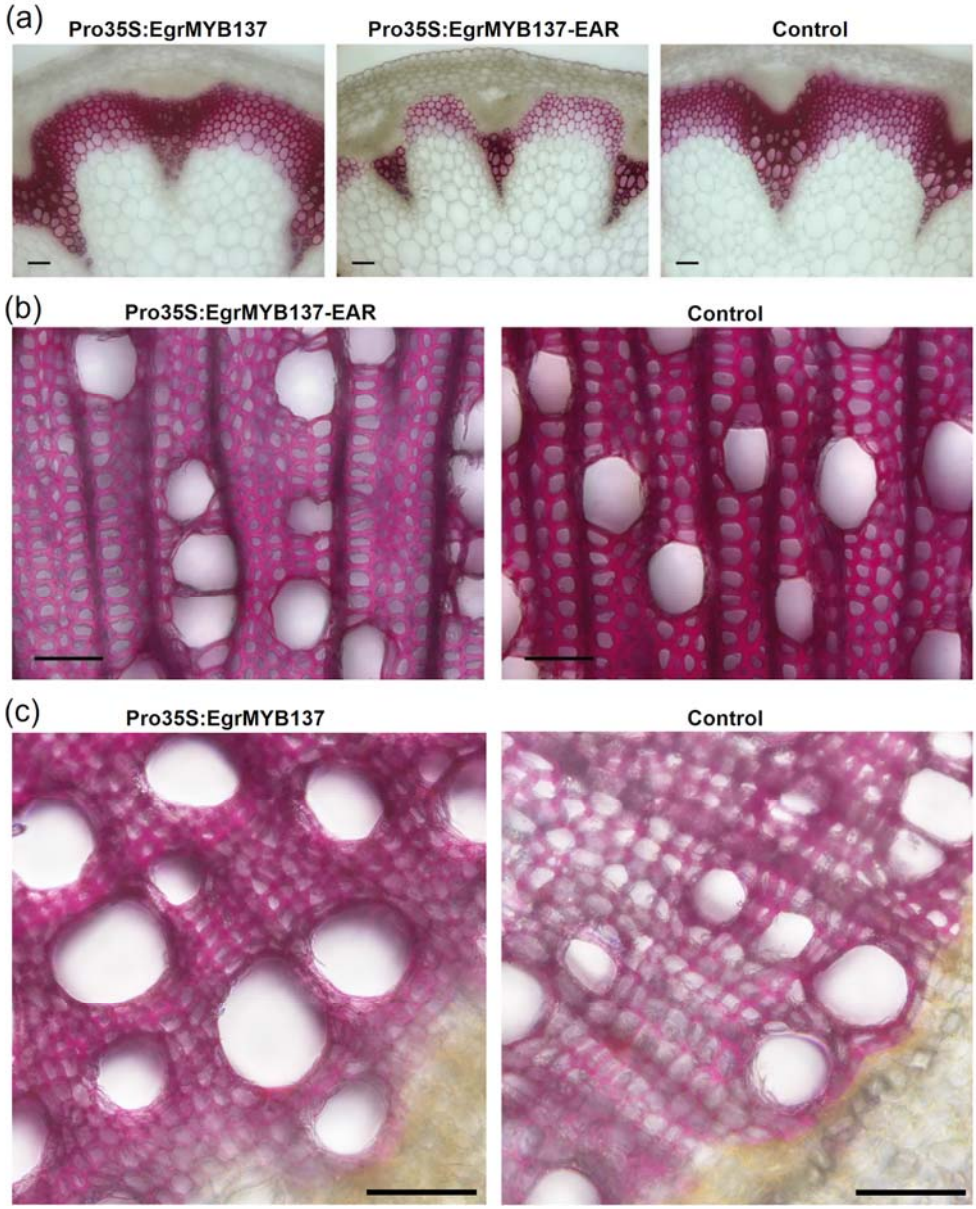


Figure 6: EgrMYB137 positively regulates xylem cell lignification in *Arabidopsis*, poplar and *Eucalyptus*. Phloroglucinol-HCl staining of transverse sections of *Arabidopsis thaliana* inflorescence stems **(a)**, *Populus tremula x alba* stems **(b)** and *Eucalyptus grandis* hairy roots **(c)** from control lines or transgenic lines (Pro35S:EgrMYB137 and Pro35S:EgrMYB137-EAR). In *Arabidopsis* and poplar, dominant-repressor lines (Pro35S:EgrMYB137-EAR) show a reduction of phloroglucinol staining while in *Eucalyptus*, overexpressing lines (Pro35S:EgrMYB137) show an augmentation of phloroglucinol staining when compared to control. Scale bar: 50 μm.

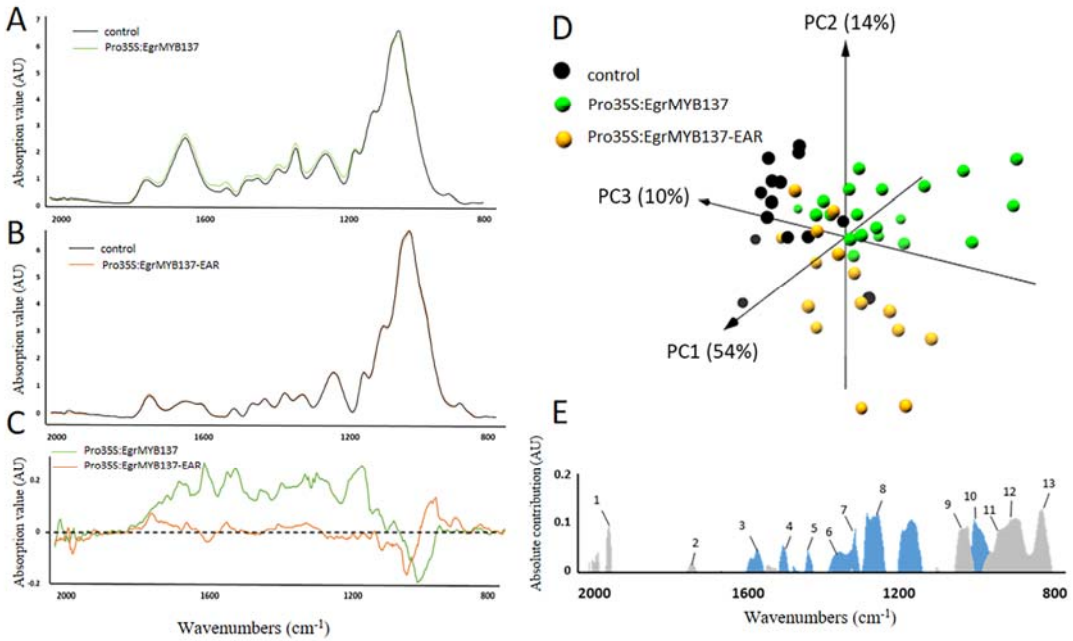
Table 2: Lignin content in EgrMYB137 overexpressing and dominant-repressor lines. Lignin content, expressed as percentage of dry weight, was measured through acetyl bromide method (AcBr) as described in Ployet et al., 2019. Measurements were performed on control lines (empty vectors), overexpressing lines (Pro35S:EgrMYB137) and dominant-repressor lines (Pro35S:EgrMYB137-EAR) obtained in 3 genetic backgrounds: *Arabidopsis thaliana*, *Populus tremula x alba* and *Eucalyptus grandis*. Asterisks indicate significant differences with control (Student's t-test; n= 4 to 9; * $p < 0.05$, *** $p < 0.001$). All measurements were done in triplicate.

	AcBr lignin content (%)
***Arabidopsis thaliana* inflorescence stems**	
Pro35S:EgrMYB137	15.2 ± 1.0
Pro35S:EgrMYB137-EAR	12.0 ± 3.1 *
Control	15.4 ± 0.6
***Populus tremula x P. alba* debarked stems**	
Pro35S:EgrMYB137-EAR	15.7 ± 1.6 ***
Control	19.5 ± 1.0
***Eucalyptus grandis* hairy roots**	
Pro35S:EgrMYB137	18.6 ± 0.4 ***
Control	16.2 ± 0.6

To gain a broader view of SCW composition in *EgrMYB137* overexpression/dominant repression transgenic lines, we performed a non-destructive analysis on SCW global composition through Fourier-Transformed Infra-Red spectroscopy (Figure 7a-c). PLS-DA performed on spectra normalised values showed a clear separation between Pro35S:EgrMYB137 lines, Pro35S:EgrMYB137-EAR lines and controls (Figure 7d), suggesting significant changes in SCW composition. A Sparse-PLS-DA analysis allowed the identification of the most discriminant wavenumbers explaining the separation between controls and transgenic lines. These wave numbers pointed to 13

regions of the spectra that were previously related to SCW components (Figure 7e). Most of these wavenumbers (12 out of 13) were associated with lignin structure and composition (Supplementary Table S13). Altogether, this evidence clearly demonstrate the positive role of EgrMYB137 in lignification.

Figure 7: Fingerprint of SCW composition in overexpressing and dominant-repressor lines using FT-IR coupled to multivariate statistical analyses. (a) FT-IR absorption spectra of controls (black) and Pro35S:EgrMYB137 *Eucalyptus* hairy root samples (green). **(b)** FT-IR absorption spectra of controls (black) and Pro35S:EgrMYB137-EAR poplar samples (orange). In (a) and (b), the curves were drawn using the median of



control and transgenic lines absorption values. **(c)** Resulting absorption spectra of *Populus* (orange) and *Eucalyptus* (green) samples when the median of the respective control is subtracted. **(d)** 3D Partial Least Square Discriminant analysis (PLS-DA) was performed using normalised spectra values of each sample (n = 14 to 20). The median spectrum of the respective control was removed from each individual spectrum (either transgenic or control

samples). The first three components explain 78% of the total variability. Black dots correspond to control samples (either poplar or *Eucalyptus* transformed with empty vectors), orange dots to Pro35S:EgrMYB137-EAR poplar samples, and green dots to Pro35S:EgrMYB137 *Eucalyptus* samples. **(e)** Identification of the 200 most discriminant wavenumbers contributing to PC1 (blue) and PC2 (grey), identified by Sparse-PLS-DA. Y axis represents the absolute value of WV loadings. Numbers indicate WN associated to cell wall components, according to data from the literature (see Supplementary Table S14).

To investigate EgrMYB137 role in SCW formation at the transcriptional level, targeted gene expression analyses were performed on Pro35S:EgrMYB137-EAR poplar lines and control lines by RT-qPCR. 53 genes involved in either SCW regulation or in SCW biosynthesis were profiled, the majority of which are close homologs of EgrMYB137 predicted direct targets. As a general trend, most of these genes are repressed in the three independent dominant repressive lines compared to control (Figure 8). Among SCW regulators, *PtVNS10*, involved in fibre differentiation, is significantly repressed as well as three MYB TFs (*PtrMYB216, PtrMYB125, PtrMYB221*). Three out of seventeen tested polysaccharides biosynthesis genes were significantly down-regulated in Pro35S:EgrMYB137-EAR poplar lines, among them one cellulose synthase (*PtrCesA8B*) and two genes involved in hemicellulose biosynthesis (*PtrGT47c and PtrGXM3*). Regarding lignin biosynthesis, nine genes out of nineteen were significantly repressed in at least one of the dominant repressive lines, all but one of which are homologs of *E. grandis* genes predicted to be direct targets of EgrMYB137 (Figure 4).

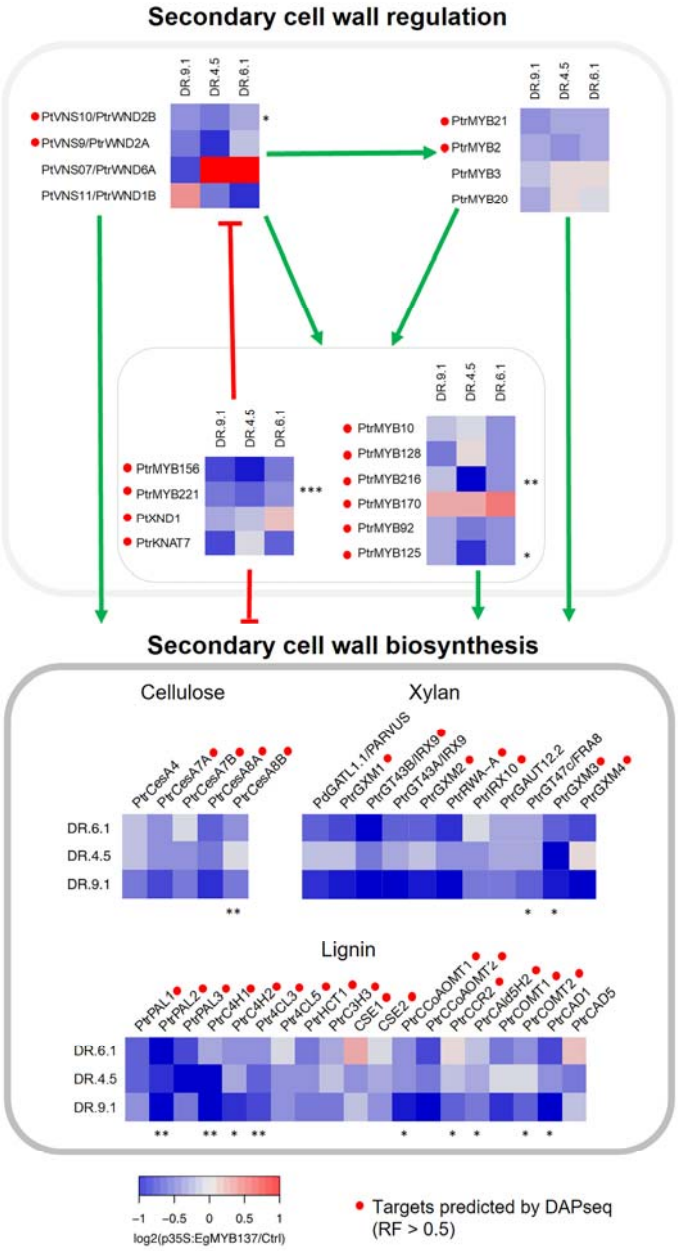


Figure 8: The dominant repressive form of EgrMYB137 represses the transcription of cell wall-related genes in poplar. Expression analyses were performed by RT-qPCR on three biological replicates of three independent Pro35S:EgrMYB137-EAR lines named DR.4.5, DR.6.1 and DR.9.1) and 4 control plants (empty vector lines). Heatmaps represent the average expression ratio between

Pro35S:EgrMYB137-EAR transgenic and control lines. The red-to-green colour scale represents log2 of expression ratios. *significant up/down-regulation between Pro35S:EgrMYB137-EAR transgenic plants and control plants: * between 1 line and controls, ** between 2 lines and control, *** between 3 lines and control (Student's T-test, n = 3, P-value <0.05). The green and red arrows represent positive and negative transcriptional regulation respectively.

Discussion

In this study, we explored the application of DAP-seq-ML to better understand the regulatory functions of EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135, and EgrMYB137, all of which have been linked to SCW regulation either directly or via their homology with SCW-related genes in other lineages. Identifying the direct gene targets of a TF is a step toward deciphering its biological function. Based on the DAP-seq-ML results for EgrMYB1, EgrMYB2, EgrMYB122 and EgrMYB137, we demonstrate that assigning predicted target genes to peaks via a random forest classifier differentiates biologically relevant (that is, SCW-related) predicted gene targets from incidental associations. To further validate this, we show using *Arabidopsis*, *Populus*, and *Eucalyptus* transgenic systems that *EgrMYB137* promotes SCW lignification *in planta* and affects the expression of SCW-related structural and regulatory genes, most of which were predicted to be direct targets using the DAP-seq-ML pipeline. We have also contributed improvements to the workflow of DAP-seq through additional quality control and target gene inference to identify TFBSs and biologically meaningful gene targets more reliably.

The pre-ML predicted target genes for EgrMYB1, EgrMYB2 and EgrMYB137 had low enrichment of the expected GO terms before the implementation of our ML classifier algorithm probability cut-offs. Having demonstrated that the ML classifier is able to improve target gene assignment in *Arabidopsis*, we provide evidence from the well-known AtMYB83 ortholog EgrMYB2, AtMYB4 ortholog EgrMYB1 as well as our functional investigation of EgrMYB137 that its application in *Eucalyptus grandis* improved the reliability of target gene identification. We present evidence of improved enrichment for SCW-related GO terms especially for EgrMYB2 (Supplementary Table S9), EgrMYB1 (Supplementary Table S10), and EgrMYB137 (Supplementary Table S11) which supports the reliability of the improved target gene assignment using ML. Additionally, EgrMYB122, phenylpropanoid biosynthesis was only enriched among targets at Pr ≥0.5, consistent with the activation of the phenylpropanoid pathway by its close homolog AtMYB43 (Geng et al. 2020) (Supplementary Table S12). Of the 11 non-redundant learning features used in our classifier, co-expression was by far the most important, while chromatin-based features (DHSs) and whether TFBS occur at conserved positions contributed little to the classifier's performance (Supplementary Figure S5). A recent study exploring lipid metabolism regulation in *Camelina sativa* showed that a high-confidence set of TFs involved in lipid metabolism regulation can be attained from DAP-seq data coupled with co-expression analysis (Gomez-Cano et al. 2022). Gene expression data is also comparatively more abundant for non-model organisms than chromatin data. A standing question for further research regarding the DAP-seq-ML approach is to what extent a model trained on features from one species is transferrable to similar datasets from another, since dissimilarities in the distributions of features between

independent domains (i.e., *Arabidopsis* vs target species) may degrade ML performance in the absence of a domain adaptation approach (Farahani et al. 2020). While the future incorporation of a limited number of *Eucalyptus*-specific labels may allow for better domain adaptation and model transferability, we argue that co-expression based on a large number of samples (hundreds to thousands, in our case) is a robust feature because we expect that a TF will, to a certain extent, be positively or negatively co-expressed with its target genes across many different tissues or conditions. While only a modest number (n = 20) of RNA-seq datasets is required to calculate co-expression metrics, there is a demonstrated preference for thousands of samples to accurately infer functional connectivity in co-expression networks (Ballouz et al. 2015). Co-expression could also potentially bias the target gene set towards SCW-related candidates given that our candidates and their orthologs are SCW-related. As evidence against this, we see that EgrMYB135 displayed no SCW-related GO terms (Supplementary Table S13) following application of the classifier.

Among the inferred gene targets for ErgMYB1, we observed several *bona fide* lignin genes such as EgrCSE, EgrCCoAOMT2, EgrC3H3, Egr4CL1, EgrCCoAOMT1, and EgrC4H1 (Carocha et al. 2015) and high enrichment for lignin-related terms. Presumably, EgrMYB1 directly represses these genes, but the nature of its interaction with linker histone EgH3.1 (Soler et al. 2016) at these loci remains to be understood. However, our target gene data is consistent with EgrMYB1's demonstrated role in inhibition of SCW deposition and repression of the lignin pathway (Legay et al. 2010; Soler et al. 2016).

GO enrichment of predicted EgrMYB122 (a homolog of AtMYB20, 42, 43 and 85) targets showed a link to response to wounding and phenylpropanoid metabolism and biosynthesis. However, among its targets was a set of *bona fide* lignin genes, known to be highly expressed in *Eucalyptus* secondary developing xylem (Carocha et al. 2015) namely, *Egr4CL1*, *EgrPAL9*, *EgrC3H3*, *EgrC4H1*, and *EgrCCoAOMT2*. Many of these have been well characterized in *Arabidopsis*. For instance, *AtPAL2*, a close ortholog of *EgrPAL9* is involved in phenylpropanoid catabolic process, defence response, phenylpropanoid biosynthetic process, response to oxidative stress and response to wounding (Cochrane et al. 2004; Rohde et al. 2004; Wong et al. 2012; Chun et al. 2019). Moreover, we also see direct targeting of EgrMYB1 (a putative ortholog of AtMYB4) as well as its paralog MYB4 (Eucgr.I01406) by EgrMYB122 (Figure 4), which is consistent with the activation of the AtMYB4 repressor by the AtMYB20/AtMYB42/ AtMYB43/AtMYB85 clade to divert phenylpropanoid flux to the lignin biosynthesis pathway through the repression of flavonoid biosynthesis (Geng et al. 2020). We similarly expected EgrMYB135 to have a strong enrichment for SCW-related GO terms, given that in *Arabidopsis* AtMYB52 (its possible ortholog), has been implicated in cell wall thickening (Zhong et al. 2008; Ko et al. 2009; Zhou et al. 2009; Romano et al. 2012; Zhang et al. 2018), the regulation of fibre cells (Zhong et al. 2008; Ko et al. 2009; Taylor-Teeples et al. 2015), abiotic response (Park et al. 2011) and the regulation of lignification (Cassan-Wang et al. 2013). While no GO terms were enriched for filtered target genes although several terms were enriched for the unfiltered sets (pre-ML and post-ML) of targets, TFs with SCW-related homologs such as *VND6*, *REV*, *KNAT7*, *MYB4*, *MYB103*, and *SND3* were among some of the targets of EgrMYB135. Possibly, EgrMYB135 could act through

one or more of these intermediate TFs to regulate SCW deposition, without SCW-related biological processes being enriched among its direct targets.

The biological role of EgrMYB2 inferred from this study, supported by the enriched GO terms and inferred target genes for EgrMYB2 (a positive SCW master regulator (Goicoechea et al. 2005; Zhong et al. 2008; McCarthy et al. 2009; Zhong et al. 2010; Zhong and Ye 2012; Zhong et al. 2013; Ko et al. 2014)) largely agrees with our expectation that EgrMYB2 is a conserved SCW master regulator. Despite some of the targets of EgrMYB2's orthologs were revealed in *Arabidopsis* and *Populus* (Ko et al. 2009; Zhong and Ye 2012; Chen et al. 2019), we report here for the first time an exhaustive set of genome-wide targets for EgrMYB2. Like AtMYB46 and AtMYB83, EgrMYB2 directly targets genes involved in hemicellulose, cellulose, and lignin regulation and a cascade of TFs that occupy different positions in the SCW-related transcriptional regulatory network (Figure 4). One third of AtMYB83 targets in *Arabidopsis* (O'Malley *et al.,* 2016) are close orthologs of EgrMYB2 targets in *Eucalyptus* (see Supplementary methods S2 and Supplementary Dataset D5). This is evidence that EgrMYB2's role in regulating the SCW program is conserved. To further support the conservation of the master regulatory role of EgrMYB2, the GO enrichment analysis shows clear enrichment of SCW-related processes as well as a wide range of other generic GO terms that are not necessarily related to SCW. Interestingly the lesser known EgrMYB137 had a statistically significant overlap of target genes with SCW master regulator EgrMYB2 and they shared similar GO terms. This implies that these two TFs, which belong to different subgroups (Soler et al. 2015), might be involved in similar biological processes and possibly co-regulate the

same biological pathways. We identified genome-wide targets of EgrMYB137, providing new insights into the role of this TF in SCW deposition. *EgrMYB137* is preferentially expressed in xylem tissues and belongs to subgroup 13 of the R2R3-MYB family, enriched in MYBs controlling SCW deposition (Soler et al. 2015). EgrMYB137 was first characterized as a regulator of xylem development in *Eucalyptus*, co-expressed with 25 SCW-related genes, including 5 TFs of SCW regulation (Ployet et al. 2019). Among the direct targets of EgrMYB137 are several TFs belonging to different SCW transcriptional network tiers (Figure 4), many of which are close orthologs of TFs regulating SCW deposition in *Arabidopsis*. The general trend of down-regulation, observed for most SCW-related TFs in poplar transgenic lines where *EgrMYB137* was expressed as a dominant repressor, provides clear evidence that this TF regulates SCW-related genes *in planta* to modulate xylem tissue development and SCW deposition. Our results suggest that EgrMYB137 acts as a positive regulator mostly dedicated to regulate SCW lignification. Almost all the genes (13 out of 17 previously described by Carocha et al. (2015) involved in 10 steps of lignin biosynthesis were identified as confident targets by DAP-seq-ML analyses and most of them (9 out of 13) were significantly downregulated in Pro35S:EgrMYB137-EAR transgenic poplar lines. The role of EgrMYB137 in regulating the phenylpropanoid pathway is also supported by the functional characterisation of transgenic lines in three different plant backgrounds. When the repressive form of EgrMYB137-EAR is expressed, lignin deposition is reduced, whereas a constitutive expression of EgrMYB137 increases xylem lignification. The reduction of vascular tissue lignification in *Arabidopsis* and *Populus* EAR lines likely explains the growth defect, leaf curvature, and lack of stem rigidity as reported for

lignification defective mutants (reviewed by Yoon et al., 2015) and El Houari et al., 2021). On the contrary, an increase in lignification was detected only in overexpressing *Eucalyptus* HR lines. This could be explained by the strong effect of the repressive EAR domain (Hiratsu et al. 2003) to destabilise SCW regulation whereas the effect of *EgrMYB137* overexpression in heterologous plant backgrounds could be affected by cis-regulatory divergence between lineages or counterbalanced by regulatory feedback loops to prevent over-lignification.

Apart from lignification, EgrMYB137 DAP-seq targets were enriched in GO terms related to polysaccharides biosynthesis, suggesting a broader role of EgrMYB137 in SCW deposition. FT-IR chemotyping of xylem tissue in *Eucalyptus* and *Populus* transgenic lines pointed out spectral modifications associated with polysaccharide compounds, in addition to many wavenumbers related to lignin. Our DAP-seq results suggest that EgrMYB137 could regulate SCW composition by targeting 14 genes involved in cellulose biosynthesis (including three cellulose synthases) and 30 genes involved in hemicelluloses biosynthesis.

Two very close poplar orthologs of EgrMYB137, PtrMYB074 (*P. tremula*) and PtoMYB074 (*P. tomentosa*), are directly linked to the regulation of xylem cells differentiation and SCW deposition (Romano et al. 2012; Li et al. 2018; Liu et al. 2022). Using differentiating xylem protoplasts transfection, Lin et al. (2013) demonstrated that PtrMYB74 was directly regulated by the tier 1 master switch SND1 (Secondary Wall-Associated NAC Domain 1). In poplar, PtrMYB074 acts as a tier 2 TF within SCW network, regulating tier 2 and 1 TFs as well as genes related to SCW compounds biosynthesis (Chen et al. 2019). More recently, Y2H and ChIP

experiments were used to dissect PtrMYB074 regulatory network. PtrMYB074 interacts with PtrWRKY19 to induce the expression of *PtrbHLH186* which in turn increase SCW lignification and increase the proportion of small vessels (Chen et al. 2019; Liu et al. 2022).

In our data, a close ortholog of PtrbHLH186 (Eucgr.C03807) is included within the unfiltered list of EgrMYB137 targets but interaction is not retained after ML filtering (Pr ≥ 0.5). As demonstrated in poplar for PtrMYB074 (Liu et al. 2022), the lack of potential partners of EgrMYB137 could explain the weak affinity towards *Eucalyptus* orthologs of PtrbHLH186 in our DAP-seq-ML. Although little evidence is available to transpose this mechanism to *Eucalyptus*, we observe that EgrMYB137 and PtrMYB074 share common targets related to the regulation of SCW synthesis and the synthesis of lignin. Analyses of target orthology showed that 40% of the targets of PtrMYB074 in *Populus* (Liu et al. 2022) are close orthologs of EgrMYB137 targets in *Eucalyptus* (see Supplementary Methods and Supplementary Dataset D5).

The functional characterization of EgrMYB137 and PtrMYB074 in transgenic plants demonstrates highly similar patterns of phenotypes related to SCW phenotypes. While overexpression of *bHLH186* and *PtoMYB074* altered xylem formation, SCW composition and vessels differentiated in a similar way to the constitutive expression of *EgrMYB137* in *Eucalyptus* hairy roots (Li et al. 2018; Ployet et al. 2019; Liu et al. 2022). Similar phenotypes are also observed in *Arabidopsis* overexpressing lines, suggesting at least a partial conservation of EgrMYB137/PtrMYB074 pathway in herbaceous dicots. To our knowledge, the

function of AtMYB86, the closest homolog of EgrMYB137, has never been investigated.

Growing evidence highlights the role of wood tissue remodelling in adaptation to environmental change. It was shown in *Arabidopsis* that several abiotic stresses can co-opt (recruit) key regulators of xylem development (Taylor-Teeples et al. 2015). Co-option of a developmental network is a means to facilitate adaptation to stress. In poplar, the direct regulation of *bHLH186* by PtrMYB074 and PtrWRKY19 leads to an increase of drought tolerance (Liu et al. 2022). In *Eucalyptus*, we showed that *EgrMYB137* is differentially expressed in response to abiotic constraints, and its expression is strongly correlated to SCW lignification, vessel size, and vessel density (Ployet et al. 2019). The role of *EgrMYB137* in cell wall lignification and xylem cells differentiation could influence sap flow inside vascular tissues, a key parameter for drought resistance (Rodriguez-Zaccaro and Groover 2019; Hadj Bachir et al. 2022). On another hand, if EgrMYB137 targets did not reveal any significant enrichment in stress GO categories, more than 200 genes likely involved in abiotic stress responses are potential targets of EgrMYB137 and need to be investigated *in planta*. This collectively suggests a broader role of EgrMYB137 and its orthologs at the crosstalk between wood tissues formation and stress responses.

In conclusion, we present an improved DAP-seq-ML workflow coupling *in vitro* binding data for five R2R3-MYB family TFs and machine learning. This yielded a set of improved confidence target genes for these five MYB-family TFs namely, EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135 and EgrMYB137 TFs, all implicated

in SCW development. Finally, we provide evidence *in planta* (functional characterization of transgenic lines in three different plant backgrounds) showing that EgrMYB137 acts as a positive regulator of SCW lignification, with a significant number of genes commonly targeted with AtMYB83 homolog EgrMYB2.

References

Ambawat S, Sharma P, Yadav NR, Yadav RC (2013) MYB transcription factor genes as regulators for plant responses: an overview. Physiol Mol Biol Plants 19:307–21. https://doi.org/10.1007/s12298-013-0179-1

Andrews S (2010) A quality control tool for high throughput sequence data. In: Babraham Bioinformatics

ARABIDOPSIS INTERACTOME MAPPING CONSORTIUM (2011) Evidence for Network Evolution in an Arabidopsis Interactome Map. Science (1979) 333:601–607

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME Suite: Tools for motif discovery and searching. Nucleic Acids Res 37. https://doi.org/10.1093/nar/gkp335

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36

Ballouz S, Verleyen W, Gillis J (2015) Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics 31:2123–2130. https://doi.org/10.1093/bioinformatics/btv118

Bargmann BOR, Marshall-Colon A, Efroni I, Ruffel S, Birnbaum KD, Coruzzi GM, Krouk G (2013) TARGET: A Transient Transformation System for Genome-Wide

Transcription Factor Target Discovery. Mol Plant 6:978–980. https://doi.org/10.1093/mp/sst010

Bar-On YM, Phillips R, Milo R (2018) The biomass distribution on Earth. Proceedings of the National Academy of Sciences 115:6506–6511. https://doi.org/10.1073/pnas.1711842115

Bartlett A, O RC, Carol Huang S, Galli M, Nery JR, Gallavotti A, Ecker JR (2017) Mapping genome-wide transcription factor binding sites using DAP-seq. Nat Protoc 12:1659–1672. https://doi.org/10.1038/nprot.2017.055

Brooks MD, Cirrone J, Pasquino A V., Alvarez JM, Swift J, Mittal S, Juang C-L, Varala K, Gutiérrez RA, Krouk G, Shasha D, Coruzzi GM (2019) Network Walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. Nat Commun 10:1569. https://doi.org/10.1038/s41467-019-09522-1

Brown K, Takawira LT, O'Neill MM, Mizrachi E, Myburg AA, Hussey SG (2019) Identification and functional evaluation of accessible chromatin associated with wood formation in Eucalyptus grandis. New Phytologist 223:1937–1951. https://doi.org/10.1111/nph.15897

Carocha V, Soler M, Hefer C, Cassan-Wang H, Fevereiro P, Myburg AA, Paiva JAP, Grima-Pettenati J (2015) Genome-wide analysis of the lignin toolbox of Eucalyptus grandis. New Phytologist 206:1297–1313. https://doi.org/10.1111/nph.13313

Cassan-Wang H, Goué N, Saidi MN, Legay S, Sivadon P, Goffner D, Grima-Pettenati J (2013) Identification of novel transcription factors regulating secondary cell wall formation in Arabidopsis. Front Plant Sci 4. https://doi.org/10.3389/fpls.2013.00189

Cassan-Wang H, Soler M, Yu H, Camargo ELO, Carocha V, Ladouce N, Savelli B, Paiva JAP, Leplé J-C, Grima-Pettenati J (2012) Reference Genes for High-Throughput Quantitative Reverse Transcription–PCR Analysis of Gene Expression in Organs and Tissues of Eucalyptus Grown in Various Environmental Conditions. Plant Cell Physiol 53:2101–2116. https://doi.org/10.1093/pcp/pcs152

Chen H, Wang JP, Liu H, Li H, Lin Y-CJ, Shi R, Yang C, Gao J, Zhou C, Li Q, Sederoff RR, Li W, Chiang VL (2019) Hierarchical Transcription Factor and Chromatin Binding Network for Wood Formation in Populus trichocarpa. Plant Cell 31:602–626. https://doi.org/10.1105/tpc.18.00620

Chun HJ, Baek D, Cho HM, Lee SH, Jin BJ, Yun D-J, Hong Y-S, Kim MC (2019) Lignin biosynthesis genes play critical roles in the adaptation of Arabidopsis plants to high-salt stress. Plant Signal Behav 14:1625697. https://doi.org/10.1080/15592324.2019.1625697

Cochrane FC, Davin LB, Lewis NG (2004) The Arabidopsis phenylalanine ammonia lyase gene family: kinetic characterization of the four PAL isoforms. Phytochemistry 65:1557–1564. https://doi.org/10.1016/j.phytochem.2004.05.006

da Silveira Falavigna V, Severing E, Lai X, Estevan J, Farrera I, Hugouvieux V, Revers LF, Zubieta C, Coupland G, Costes E, Andrés F (2021) Unraveling the role of MADS transcription factor complexes in apple tree dormancy. New Phytologist 232:2071–2088. https://doi.org/10.1111/nph.17710

Ding A, Tang X, Yang D, Wang M, Ren A, Xu Z, Hu R, Zhou G, O'Neill M, Kong Y (2021) ERF4 and MYB52 transcription factors play antagonistic roles in regulating homogalacturonan de-methylesterification in Arabidopsis seed coat mucilage. Plant Cell 33:381–403. https://doi.org/10.1093/plcell/koaa031

Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L (2010) MYB transcription factors in Arabidopsis. Trends Plant Sci 15:573–581. https://doi.org/10.1016/j.tplants.2010.06.005

el Houari I, Boerjan W, Vanholme B (2021) Behind the Scenes: The Impact of Bioactive Phenylpropanoids on the Growth Phenotypes of Arabidopsis Lignin Mutants. Front Plant Sci 12. https://doi.org/10.3389/fpls.2021.734070

Farahani A, Voghoei S, Rasheed K, Arabnia HR (2020) A Brief Review of Domain Adaptation

Fornalé S, Shi X, Chai C, Encina A, Irar S, Capellades M, Fuguet E, Torres JL, Rovira P, Puigdomènech P, Rigau J, Grotewold E, Gray J, Caparrós-Ruiz D (2010) ZmMYB31 directly represses maize lignin genes and redirects the phenylpropanoid metabolic flux. Plant Journal 64:633–644. https://doi.org/10.1111/j.1365-313X.2010.04363.x

Galli M, Khakhar A, Lu Z, Chen Z, Sen S, Joshi T, Nemhauser JL, Schmitz RJ, Gallavotti A (2018) The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. Nat Commun 9. https://doi.org/10.1038/s41467-018-06977-6

Geng P, Zhang S, Liu J, Zhao C, Wu J, Cao Y, Fu C, Han X, He H, Zhao Q (2020) MYB20, MYB42, MYB43, and MYB85 regulate phenylalanine and lignin biosynthesis during secondary cell wall formation1[OPEN]. Plant Physiol 182:1272–1283. https://doi.org/10.1104/PP.19.01070

Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N, Grima-Pettenati J (2005) EgMYB2, a new transcriptional activator from Eucalyptus xylem, regulates secondary cell wall formation and lignin biosynthesis. Plant Journal 43:553–567. https://doi.org/10.1111/j.1365-313X.2005.02480.x

Gomez-Cano F, Chu Y, Cruz-Gomez M, Abdullah HM, Lee YS, Schnell DJ, Grotewold E (2022) Exploring Camelina sativa lipid metabolism regulation by combining gene co-expression and DNA affinity purification analyses. The Plant Journal 110:589–606. https://doi.org/10.1111/tpj.15682

Guo Y, Mahony S, Gifford DK (2012) High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. PLoS Comput Biol 8:e1002638. https://doi.org/10.1371/journal.pcbi.1002638

Hadj Bachir I, Ployet R, Teulières C, Cassan-Wang H, Mounet F, Grima-Pettenati J (2022) Regulation of secondary cell wall lignification by abiotic and biotic constraints. Elsevier

Han Z, Yang T, Guo Y, Cui WH, Yao LJ, Li G, Wu AM, Li JH, Liu LJ (2021) The transcription factor PagLBD3 contributes to the regulation of secondary growth in Populus. J Exp Bot 72:7092–7106. https://doi.org/10.1093/jxb/erab351

Hatton D, Sablowski R, Yung M-H, Smith C, Schuch W, Bevan M (1995) Two classes of cis sequences contribute to tissue-specific expression of a PAL2 promoter in transgenic tobacco. The Plant Journal 7:859–876. https://doi.org/10.1046/j.1365-313X.1995.07060859.x

Hiratsu K, Matsui K, Koyama T, Ohme-Takagi M (2003) Dominant repression of target genes by chimeric repressors that include the EAR motif, a repression domain, in Arabidopsis. The Plant Journal 34:733–739. https://doi.org/10.1046/j.1365-313X.2003.01759.x

Houston K, Tucker MR, Chowdhury J, Shirley N, Little A (2016) The Plant Cell Wall: A Complex and Dynamic Structure As Revealed by the Responses of Genes under Stress Conditions. Front Plant Sci 7. https://doi.org/10.3389/fpls.2016.00984

Hussey SG (2022) Transcriptional regulation of secondary cell wall formation and lignification. In: Advances in Botanical Research. Academic Press Inc.

Hussey SG, Saïdi MN, Hefer CA, Myburg AA, Grima-Pettenati J (2015) Structural, evolutionary and functional analysis of the NAC domain protein family in Eucalyptus. New Phytologist 206:1337–1350. https://doi.org/10.1111/nph.13139

Jin H, Cominelli E, Bailey P, Parr A, Mehrtens F, Jones J, Tonelli C, Weisshaar B, Martin C (2000) Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. EMBO J 19:6150–6161. https://doi.org/10.1093/emboj/19.22.6150

Jin H, Martin C (1999) Multifunctionality and diversity within the plant MYB-gene family. Plant Mol Biol 41:577–585. https://doi.org/https://doi.org/10.1023/a:1006319732410

Kaufmann K, Muiño JM, Jauregui R, Airoldi CA, Smaczniak C, Krajewski P, Angenent GC (2009) Target Genes of the MADS Transcription Factor SEPALLATA3: Integration of Developmental and Hormonal Pathways in the Arabidopsis Flower. PLoS Biol 7:e1000090. https://doi.org/10.1371/journal.pbio.1000090

Ko JH, Jeon HW, Kim WC, Kim JY, Han KH (2014) The MYB46/MYB83-mediated transcriptional regulatory programme is a gatekeeper of secondary wall biosynthesis. Ann Bot 114:1099–1107

Ko J-H, Kim W-C, Han K-H (2009) Ectopic expression of MYB46 identifies transcriptional regulatory genes involved in secondary wall biosynthesis in

Arabidopsis. Plant J 60:649–665. https://doi.org/10.1111/j.1365-313X.2009.03989.x

Koncz C, Schell J (1986) The promoter of TL-DNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of Agrobacterium binary vector. Mol Gen Genet 204:383–396. https://doi.org/10.1007/BF00331014

Kumar M, Campbell L, Turner S (2016) Secondary cell walls: biosynthesis and manipulation. J Exp Bot 67:515–531. https://doi.org/10.1093/jxb/erv533

Lai X, Stigliani A, Lucas J, Hugouvieux V, Parcy F, Zubieta C (2020) Genome-wide binding of SEPALLATA3 and AGAMOUS complexes determined by sequential DNA-affinity purification sequencing. Nucleic Acids Res 48:9637–9648. https://doi.org/10.1093/nar/gkaa729

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko P V., Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813–1831. https://doi.org/10.1101/gr.136184.111

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10. https://doi.org/10.1186/gb-2009-10-3-r25

Laubscher M, Brown K, Tonfack LB, Myburg AA, Mizrachi E, Hussey SG (2018) Temporal analysis of Arabidopsis genes activated by Eucalyptus grandis NAC transcription factors associated with xylem fibre and vessel development. Sci Rep 8:10983. https://doi.org/10.1038/s41598-018-29278-w

Legay S, Lacombe E, Goicoechea M, Brière C, Séguin A, Mackay J, Grima-Pettenati J (2007) Molecular characterization of EgMYB1, a putative transcriptional repressor of the lignin biosynthetic pathway. Plant Science 173:542–549. https://doi.org/10.1016/j.plantsci.2007.08.007

Legay S, Sivadon P, Blervacq A, Pavy N, Baghdady A, Tremblay L, Levasseur C, Ladouce N, Lapierre C, Séguin A, Hawkins S, Mackay J, Grima-Pettenati J (2010) EgMYB1 , an R2R3 MYB transcription factor from eucalyptus negatively regulates secondary cell wall formation in Arabidopsis and poplar. New Phytologist 188:774–786. https://doi.org/10.1111/j.1469-8137.2010.03432.x

Li C, Ma X, Yu H, Fu Y, Luo K (2018) Ectopic Expression of PtoMYB74 in Poplar and Arabidopsis Promotes Secondary Cell Wall Formation. Front Plant Sci 9. https://doi.org/10.3389/fpls.2018.01262

Li Q, Brown JB, Huang H, Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. Ann Appl Stat 5. https://doi.org/10.1214/11-AOAS466

Lin Y-C, Li W, Sun Y-H, Kumari S, Wei H, Li Q, Tunlaya-Anukit S, Sederoff RR, Chiang VL (2013) SND1 Transcription Factor-Directed Quantitative Functional Hierarchical Genetic Regulatory Network in Wood Formation in Populus trichocarpa. Plant Cell 25:4324–4341. https://doi.org/10.1105/tpc.113.117697

Liu B, Liu J, Yu J, Wang Z, Sun Y, Li S, Lin Y-CJ, Chiang VL, Li W, Wang JP (2021) Transcriptional reprogramming of xylem cell wall biosynthesis in tension wood. Plant Physiol 186:250–269. https://doi.org/10.1093/plphys/kiab038

Liu H, Gao J, Sun J, Li S, Zhang B, Wang Z, Zhou C, Sulis DB, Wang JP, Chiang VL, Li W (2022) Dimerization of PtrMYB074 and PtrWRKY19 mediates transcriptional activation of PtrbHLH186 for secondary xylem development in Populus trichocarpa. New Phytologist 234:918–933. https://doi.org/10.1111/nph.18028

Luo L, Li L (2022) Molecular understanding of wood formation in trees. Forestry Research 2:1–11. https://doi.org/10.48130/FR-2022-0005

Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21:3448–3449. https://doi.org/10.1093/bioinformatics/bti551

McCarthy RL, Zhong R, Ye ZH (2009) MYB83 is a direct target of SND1 and acts redundantly with MYB46 in the regulation of secondary cell wall biosynthesis in arabidopsis. Plant Cell Physiol 50:1950–1964. https://doi.org/10.1093/pcp/pcp139

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, Goodstein DM, Dubchak I, Poliakov A, Mizrachi E, Kullan ARK, Hussey SG, Pinard D, Van Der Merwe K, Singh P, Van Jaarsveld I, Silva-Junior OB, Togawa RC, Pappas MR, Faria DA, Sansaloni CP, Petroli CD, Yang X, Ranjan P, Tschaplinski TJ, Ye CY, Li T, Sterck L, Vanneste K, Murat F, Soler M, Clemente HS, Saidi N, Cassan-Wang H, Dunand C, Hefer CA, Bornberg-Bauer E, Kersting AR, Vining K, Amarasinghe V, Ranik M, Naithani S, Elser J, Boyd AE, Liston A, Spatafora JW, Dharmwardhana P, Raja R, Sullivan C, Romanel E, Alves-Ferreira M, Külheim C, Foley W, Carocha V, Paiva J, Kudrna D, Brommonschenkel SH,

Pasquali G, Byrne M, Rigault P, Tibbits J, Spokevicius A, Jones RC, Steane DA, Vaillancourt RE, Potts BM, Joubert F, Barry K, Pappas GJ, Strauss SH, Jaiswal P, Grima-Pettenati J, Salse J, Van De Peer Y, Rokhsar DS, Schmutz J (2014) The genome of Eucalyptus grandis. Nature 510:356–362. https://doi.org/10.1038/nature13308

Nakano Y, Yamaguchi M, Endo H, Rejab NA, Ohtani M (2015) NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. Front Plant Sci 6. https://doi.org/10.3389/fpls.2015.00288

O'Malley RC, Huang SSC, Song L, Lewsey MG, Bartlett A, Nery JR, Galli M, Gallavotti A, Ecker JR (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell 165:1280–1292. https://doi.org/10.1016/j.cell.2016.04.038

Park MY, Kang J, Kim SY (2011) Overexpression of AtMYB52 confers ABA hypersensitivity and drought tolerance. Mol Cells 31:447–454. https://doi.org/10.1007/s10059-011-0300-7

Patzlaff A, McInnis S, Courtenay A, Surman C, Newman LJ, Smith C, Bevan MW, Mansfield S, Whetten RW, Sederoff RR, Campbell MM (2003) Characterisation of a pine MYB that regulates lignification. Plant Journal 36:743–754. https://doi.org/10.1046/j.1365-313X.2003.01916.x

Pitre FE, Pollet B, Lafarguette F, Cooke JEK, MacKay JJ, Lapierre C (2007) Effects of Increased Nitrogen Supply on the Lignification of Poplar Wood. J Agric Food Chem 55:10306–10314. https://doi.org/10.1021/jf071611e

Plasencia A, Soler M, Dupas A, Ladouce N, Silva-Martins G, Martinez Y, Lapierre C, Franche C, Truchet I, Grima-Pettenati J (2016) Eucalyptus hairy roots, a fast, efficient and versatile tool to explore function and expression of genes involved in

wood formation. Plant Biotechnol J 14:1381–1393. https://doi.org/10.1111/pbi.12502

Ployet R, Veneziano Labate MT, Regiani Cataldi T, Christina M, Morel M, San Clemente H, Denis M, Favreau B, Tomazello Filho M, Laclau J, Labate CA, Chaix G, Grima-Pettenati J, Mounet F (2019) A systems biology view of wood formation in Eucalyptus grandis trees submitted to different potassium and water regimes. New Phytologist 223:766–782. https://doi.org/10.1111/nph.15802

Qu G, Peng D, Yu Z, Chen X, Cheng X, Yang Y, Ye T, Lv Q, Ji W, Deng X, Zhou B (2021) Advances in the role of auxin for transcriptional regulation of lignin biosynthesis. Functional Plant Biology 48:743. https://doi.org/10.1071/FP20381

Raes J, Rohde A, Christensen JH, van de Peer Y, Boerjan W (2003) Genome-Wide Characterization of the Lignification Toolbox in Arabidopsis . Plant Physiol 133:1051–1071. https://doi.org/10.1104/pp.103.026484

Rodriguez-Zaccaro FD, Groover A (2019) Wood and water: How trees modify wood development to cope with drought. PLANTS, PEOPLE, PLANET 1:346–355. https://doi.org/10.1002/ppp3.29

Rohde A, Morreel K, Ralph J, Goeminne G, Hostyn V, de Rycke R, Kushnir S, van Doorsselaere J, Joseleau J-P, Vuylsteke M, van Driessche G, van Beeumen J, Messens E, Boerjan W (2004) Molecular Phenotyping of the pal1 and pal2 Mutants of Arabidopsis thaliana Reveals Far-Reaching Consequences on Phenylpropanoid, Amino Acid, and Carbohydrate Metabolism. Plant Cell 16:2749–2771. https://doi.org/10.1105/tpc.104.023705

Romano JM, Dubos C, Prouse MB, Wilkins O, Hong H, Poole M, Kang K, Li E, Douglas CJ, Western TL, Mansfield SD, Campbell MM (2012) AtMYB61, an R2R3-MYB

transcription factor, functions as a pleiotropic regulator via a small gene network. New Phytologist 195:774–786. https://doi.org/10.1111/j.1469-8137.2012.04201.x

Roy S (2016) Function of MYB domain transcription factors in abiotic stress and epigenetic control of stress response in plant genome. Plant Signal Behav 11:e1117723. https://doi.org/10.1080/15592324.2015.1117723

Shen H, He X, Poovaiah CR, Wuddineh WA, Ma J, Mann DGJ, Wang H, Jackson L, Tang Y, Neal Stewart J, Chen F, Dixon RA (2012) Functional characterization of the switchgrass (Panicum virgatum) R2R3-MYB transcription factor PvMYB4 for improvement of lignocellulosic feedstocks. New Phytologist 193:121–136. https://doi.org/10.1111/j.1469-8137.2011.03922.x

Shi D, Ren A, Tang X, Qi G, Xu Z, Chai G, Hu R, Zhou G, Kong Y (2018) MYB52 Negatively Regulates Pectin Demethylesterification in Seed Coat Mucilage. Plant Physiol 176:2737–2749. https://doi.org/10.1104/pp.17.01771

Shimada TL, Shimada T, Hara-Nishimura I (2010) A rapid and non-destructive screenable marker, FAST, for identifying transformed seeds of Arabidopsis thaliana. The Plant Journal 61:519–528. https://doi.org/10.1111/j.1365-313X.2009.04060.x

Sixto H, González-González BD, Molina-Rueda JJ, Garrido-Aranda A, Sanchez MM, López G, Gallardo F, Cañellas I, Mounet F, Grima-Pettenati J, Cantón F (2016) Eucalyptus spp. and Populus spp. coping with salinity stress: an approach on growth, physiological and molecular features in the context of short rotation coppice (SRC). Trees 30:1873–1891. https://doi.org/10.1007/s00468-016-1420-7

Soler M, Camargo ELO, Carocha V, Cassan-Wang H, San Clemente H, Savelli B, Hefer CA, Paiva JAP, Myburg AA, Grima-Pettenati J (2015) The Eucalyptus grandis R2R3-MYB

transcription factor family: Evidence for woody growth-related evolution and function. New Phytologist 206:1364–1377. https://doi.org/10.1111/nph.13039

Soler M, Plasencia A, Larbat R, Pouzet C, Jauneau A, Rivas S, Pesquet E, Lapierre C, Truchet I, Grima-Pettenati J (2017) The Eucalyptus linker histone variant EgH1.3 cooperates with the transcription factor EgMYB1 to control lignin biosynthesis during wood formation. New Phytologist 213:287–299. https://doi.org/10.1111/nph.14129

Soler M, Plasencia A, Lepikson-Neto J, Camargo ELO, Dupas A, Ladouce N, Pesquet E, Mounet F, Larbat R, Grima-Pettenati J (2016) The Woody-Preferential Gene EgMYB88 Regulates the Biosynthesis of Phenylpropanoid-Derived Compounds in Wood. Front Plant Sci 7. https://doi.org/10.3389/fpls.2016.01422

Song Q, Lee J, Akter S, Rogers M, Grene R, Li S (2020) Prediction of condition-specific regulatory genes using machine learning. Nucleic Acids Res 48:e62–e62. https://doi.org/10.1093/nar/gkaa264

Sullivan AM, Bubb KL, Sandstrom R, Stamatoyannopoulos JA, Queitsch C (2015) DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. Curr Plant Biol 3–4:40–47. https://doi.org/10.1016/j.cpb.2015.10.001

Sundell D, Street NR, Kumar M, Mellerowicz EJ, Kucukoglu M, Johnsson C, Kumar V, Mannapperuma C, Delhomme N, Nilsson O, Tuominen H, Pesquet E, Fischer U, Niittylä T, Sundberg B, Hvidsten TR (2017) AspWood: High-Spatial-Resolution Transcriptome Profiles Reveal Uncharacterized Modularity of Wood Formation in *Populus tremula*. Plant Cell 29:1585–1604. https://doi.org/10.1105/tpc.17.00153

Taylor-Teeples M, Lin L, De Lucas M, Turco G, Toal TW, Gaudinier A, Young NF, Trabucco GM, Veling MT, Lamothe R, Handakumbura PP, Xiong G, Wang C, Corwin J, Tsoukalas A, Zhang L, Ware D, Pauly M, Kliebenstein DJ, Dehesh K, Tagkopoulos

I, Breton G, Pruneda-Paz JL, Ahnert SE, Kay SA, Hazen SP, Brady SM (2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature 517:571–575. https://doi.org/10.1038/nature14099

van de Velde J, van Bel M, Vaneechoutte D, Vandepoele K (2016) A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants. Plant Physiol 171:2586–2598. https://doi.org/10.1104/pp.16.00821

Vimolmangkang S, Han Y, Wei G, Korban SS (2013) An apple MYB transcription factor, MdMYB3, is involved in regulation of anthocyanin biosynthesis and flower development. BMC Plant Biol 13. https://doi.org/10.1186/1471-2229-13-176

Wang X, Niu Q-W, Teng C, Li C, Mu J, Chua N-H, Zuo J (2009) Overexpression of PGA37/MYB118 and MYB115 promotes vegetative-to-embryonic transition in Arabidopsis. Cell Res 19:224–235. https://doi.org/10.1038/cr.2008.276

Wang XC, Wu J, Guan ML, Zhao CH, Geng P, Zhao Q (2020a) Arabidopsis MYB4 plays dual roles in flavonoid biosynthesis. The Plant Journal 101:637–652. https://doi.org/10.1111/tpj.14570

Wang Z, Mao Y, Guo Y, Gao J, Liu X, Li S, Lin Y-CJ, Chen H, Wang JP, Chiang VL, Li W (2020b) MYB Transcription Factor161 Mediates Feedback Regulation of Secondary wall-associated NAC-Domain1 Family Genes for Wood Formation. Plant Physiol 184:1389–1406. https://doi.org/10.1104/pp.20.01033

Wong JH, Namasivayam P, Abdullah MP (2012) The PAL2 promoter activities in relation to structural development and adaptation in Arabidopsis thaliana. Planta 235:267–277. https://doi.org/10.1007/s00425-011-1506-9

Xiao R, Zhang C, Guo X, Li H, Lu H (2021) MYB Transcription Factors and Its Regulation in Secondary Cell Wall Formation and Lignin Biosynthesis during Xylem Development. Int J Mol Sci 22:3560. https://doi.org/10.3390/ijms22073560

Xu Y, Wang Y, Wang X, Pei S, Kong Y, Hu R, Zhou G (2020) Transcription Factors BLH2 and BLH4 Regulate Demethylesterification of Homogalacturonan in Seed Mucilage. Plant Physiol 183:96–111. https://doi.org/10.1104/pp.20.00011

Yoon J, Choi H, An G (2015) Roles of lignin biosynthesis and regulatory genes in plant development. J Integr Plant Biol 57:902–912. https://doi.org/10.1111/jipb.12422

Zhang J, Xie M, Tuskan GA, Muchero W, Chen JG (2018) Recent advances in the transcriptional regulation of secondary cell wall biosynthesis in the woody plants. Front Plant Sci 871

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9:R137. https://doi.org/10.1186/gb-2008-9-9-r137

Zhong R, Lee C, Ye Z-H (2010) Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. Trends Plant Sci 15:625–632. https://doi.org/10.1016/j.tplants.2010.08.007

Zhong R, Lee C, Zhou J, McCarthy RL, Ye ZH (2008) A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis. Plant Cell 20:2763–2782. https://doi.org/10.1105/tpc.108.061325

Zhong R, McCarthy RL, Haghighat M, Ye ZH (2013) The Poplar MYB Master Switches Bind to the SMRE Site and Activate the Secondary Wall Biosynthetic Program during Wood Formation. PLoS One 8. https://doi.org/10.1371/journal.pone.0069219

Zhong R, Ye ZH (2015) Secondary cell walls: biosynthesis, patterned deposition and transcriptional regulation. Plant Cell Physiol 56:195–214

Zhong R, Ye ZH (2012) MYB46 and MYB83 bind to the SMRE sites and directly activate a suite of transcription factors and secondary wall biosynthetic genes. Plant Cell Physiol 53:368–380. https://doi.org/10.1093/pcp/pcr185

Zhong R, Ye Z-H (2007) Regulation of cell wall biosynthesis. Curr Opin Plant Biol 10:564–572. https://doi.org/10.1016/j.pbi.2007.09.001

Zhou J, Lee C, Zhong R, Ye ZH (2009) MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. Plant Cell 21:248–266. https://doi.org/10.1105/tpc.108.063321

Zhu C-C, Wang C-X, Lu C-Y, Wang J-D, Zhou Y, Xiong M, Zhang C-Q, Liu Q-Q, Li Q-F, Wang C-Y;, Zhou J-D;, Xiong Y;, Zhang M;, Liu C-Q;, Li Q-Q; (2021) Genome-Wide Identification and Expression Analysis of OsbZIP09 Target Genes in Rice Reveal Its Mechanism of Controlling Seed Germination. Int J Mol Sci 22. https://doi.org/10.3390/ijms

Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data

Statements & Declarations

Funding

Funding was provided by the National Research Foundation (NRF) (UID 129155), the Department of Science and Innovation, Mondi Ltd and Sappi Ltd. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. This work was supported by the University of Toulouse, the Centre National pour la Recherche Scientifique (CNRS) and the French Laboratory of Excellence project 'TULIP' (ANR-10-LABX-41; ANR-11- IDEX-0002-02).

Author Information & Contributions

LTT and IHB drafted the manuscript together with SGH and FM, who conceived of the study. LTT conducted DAP-seq experiments for Laboratory 1, while IHB and NL generated EgrMYB2 and EgrMYB137 DAP-seq data (Laboratory 2). LTT, IHB and HSC performed bioinformatic analytical pipelines. JT developed the machine learning algorithm under supervision of SH, NC and RP. IHB, AD, NL and RP conducted the reverse genetics work under the supervision of FM and JG-P. AR assessed the native transcriptional activation activity of EgrMYB137 in yeast under the supervision of SH and FM. EM, AAM and JG-P co-supervised the study.

OrcID numbers (where available):

Lazarus T. Takawira 0000-0001-9007-094X

Ines Hadj Bachir 0000-0002-1652-6184

Raphaël Ployet 0000-0003-4809-5988

Jade Tulloch - NA

Helene San Clemente 0000-0002-1864-6577

Nanette Christie 0000-0003-4331-2103

Nathalie Ladouce 0000-0003-2177-8084

Annabelle Dupas 0000-0003-0295-4599

Avanish Rai 0000-0002-2902-8838

Jacqueline Grima-Pettenati 0000-0003-4272-9767

Eshchar Mizrachi 0000-0002-3126-9593

Alexander A. Myburg 0000-0003-0644-5003

Fabien Mounet 0000-0001-6632-3070

Steven G. Hussey 0000-0002-1079-8373

Acknowledgments

We thank Jonathan Featherston and Dirk Swanevelder from the Agricultural Research Council, Pretoria, and Dr Tuan Dong (University of Pretoria) for their assistance with Covaris sonication. The authors acknowledge the Genotoul GeT Platform for genomic analyses and the TRI-Genotoul platform for microscopic analyses. RP and IHB were supported by a PhD grant from the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche. The authors acknowledge Marcal Soler and Anna Plasencia Casadeval for their involvement in this work.

Data availability

1. Sequencing data were registered at the NCBI SRA database (PRJNA899014 and PRJNA899708).
2. Supplementary Datasets
    a. Dataset D1: List of *Arabidopsis* RNA-seq datasets considered in this study.
    b. Dataset D2: List of *Eucalyptus* RNA-seq datasets considered in this study.
    c. Dataset D3: Direct and indirect protein-DNA interactions and protein-protein interactions involved in Arabidopsis secondary cell wall transcriptional regulation.
    d. Dataset D4: List of 147 candidate Eucalyptus secondary cell wall-related transcription factor genes
    e. Dataset D5: DAP-seq-ML gene targets and associated peak sets for EgrMYB1, EgrMY2, EgrMYB122, EgrMYB135 and EgrMYB137.

Ethics declaration

Ethical approval for study was obtained at the University of Pretoria (protocol number NAS105/2021).

Supplementary information

1. Supplementary Information
    a. Supplementary Tables S1 - S14

b. Supplementary Figures S1 - S5

2. Supplementary Methods
    a. S1- Genomic DNA library preparation & DNA Binding assays
    b. S2- Machine learning classifier for target gene identification & Identification of SCW-associated transcription factors and structural genes in *E. grandis*
    c. S3- FT-IR analysis