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

Supplementary Methods S1

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. Genomic DNA was extracted using the NucleoSpin Plant Kit (Machery-Nagel). Two minimal adapter-ligated genomic DNA template samples (herein called E1 and E2) were independently prepared for DAP-seq analysis at the University of Pretoria. Additionally, *E. grandis* gDNA was shipped to the University of Toulouse for independent sonication, adapter-ligated library preparation and DAP-seq analysis of EgrMYB1 and EgrMYB2 (E3) or EgrMYB137 (E4). Genomic DNA library preparation, DNA affinity purification, library amplification (introduction of indexes) and library pooling were all carried out as previously described with minor modifications (Bartlett et al. 2017). For protein binding, 300 µl of Magne ® HaloTag ® Beads was transferred to a clean 1.5 mL tube instead of the prescribed 1 mL. Additionally, a purified HaloTag® Standard Protein (Cat. # G4491) (Promega, Madison, USA) or unfused HaloTag expressed from an empty pIX-HALO plasmid was introduced as a negative binding control. 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 (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). qPCRs were performed using 1 µl of DNA templates (2 ng/µl) and non-ligated gDNA template as negative control. We used A / B primers as described in Bartlett et al. (2017). SPRIselect bead-based Double Size Selection (Beckman Coulter,

Brea, CA, USA) was performed on pooled DNA libraries as 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 pool 3 independent DAP experiments per TF. To achieve an equimolar pooling of replicate libraries, we compared two methodologies. For the first one, nucleic acid concentration was measured classically by spectroscopy. For the second one, we measured exclusively adaptor-ligated gDNA concentration through the integration of the 320 bp peaks area by capillary electrophoresis software (Agilent©).

Supplementary Methods S2

Machine learning classifier for target gene identification

Fourteen TFs with both DAP-seq data (O'Malley et al. 2016) and TF perturbation data following transient induction in protoplasts in the presence of cyloheximide (Brooks et al. 2019) were used to obtain TFBS-gene associations and true target gene labels (that is, those with evidence of differential expression following perturbation of the TF). Either the ampDAP-seq or DAP-seq datasets with the best percentage of reads in peaks (FRiP) for each TF, as well as their motif positional weight matrices (PWMs) were downloaded from the Plant Cistrome Database (O'Malley et al. 2016). Potential gene targets were assigned to the closest DAP-seq peak for each TF using ChIPpeakAnno (Zhu et al. 2010) if it fell within 5 kb of the transcription start site (TSS). From these low-confidence TF-gene associations, 13,620 positive learning examples were identified as those that were differentially expressed in root cells in response to perturbation of the corresponding TF (Brooks et al. 2019). Three sets of negative learning examples were selected by considering genes that were not differentially expressed in root cells in response to TF perturbation: (i) undetected genes (UDGs; n = 14,745) with a value of zero TPM in root tissue (Brooks et al. 2019; NCBI GEO database accession GSE117857); (ii) lowly expressed genes (LEGs; n = 11,971) with TPM between 0 and 5 in root tissue and (iii) a random subset of genes (RANDOMs; n = 13,620) regardless of their expression levels.

A feature matrix was constructed from each negative sample set, using a balanced set up (similar number of positive and negative training examples). Twenty-two features were extracted 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 (**Table S2**). For the latter, 2,479 transcriptomes from 33 published *Arabidopsis* experiments (Supplementary dataset D1) were standardised by re-mapping the publicly available raw data and calculating TPM gene expression values for calculating Pearson and Spearman correlation coefficients between each TF-target candidate.

Raw sequencing files were retrieved from the online repository SRA (https://www.ncbi.nlm.nih.gov/sra) using the prefetch function available in the SRA toolkit (v2.10.0). All transcriptomes were pre-processed through a pipeline similar to Sundell et al. (2017), involving four main steps: (i) quality controls of the reads, ii) trimming of the reads, iii) mapping of the reads to reference genome and iv) gene expression quantification. Quality controls were performed using FastQC (v0.11.5; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality bases and Illumina adapters were trimmed using Trimmomatic v0.36 (illuminaclip seed mismatches: 2; palindrome clip threshold: 30; simple clip threshold: 10; headcrop: 10; avgqual: 20) (Bolger et al. 2014). After another round of quality controls using FastQC, reads were mapped on the reference genome (Arabidopsis TAIR10) retrieved from Phytozome v12 (Goodstein et al. 2012). Mapping was performed with STAR (Dobin et al. 2013) excluding reads mapping at multiple positions (outFilterMultimapNmax 1). Gene expression (TPM) was then quantified using

Stringtie (Pertea et al. 2015) by considering only reads mapping in regions annotated as genes in the genome. Quality controls were then performed on TPM values in R. Briefly, for each experiment, outlier samples were detected through hierarchical clustering and principal component analysis (PCA) using 'mixomics' R library (Rohart et al. 2017), and excluded from the final expression matrix.

Features relating to the enrichment and significance of the DAP-seq peaks were extracted from the narrowPeak or ChIPpeakAnno output files. Logistic regression (LR), support vector machine (SVM), and random forest (RF) classifiers were trained, tested, and evaluated for classifying the TF-gene associations as true or false. Each feature matrix was split into independent training (80%) and testing (20%) subsets and feature values transformed by mean standardization. Five-fold cross-validation of the training data using the area under the receiver operating characteristic curve (AUC-ROC) as a performance metric was used to optimise the model parameters, after which test set labels were predicted. For model implementation in *Eucalyptus*, an 11-feature matrix was constructed for similar categories of *E. grandis* data for TFs EgrMYB1, EgrMYB2, EgrMYB122, EgrMYB135 and EgrMYB137 using the DAP-seq and motif PWM data reported in this study as well as published DNase-seq (Brown et al. 2019) and CNS data (van de Velde et al. 2016) for *Eucalyptus*. We similarly constructed a TPM matrix of 651 published RNA-seq samples to calculate the co-expression feature (Supplementary Dataset D2). A similar pipeline as for *Arabidopsis* data was used for reprocessing RNA-seq data from *Eucalyptus*, with mapping to the latest reference genome (*Eucalyptus* V2.0) retrieved from Phytozome v12. Potential TF-target pairs were identified by assigning

DAP-seq peaks to gene models using ChIPpeakAnno as before (based on a 5 kb binding region around the gene TSS), and the predicted target genes (here defined as having a probability ≥ 0.5 from the classifier output) extracted from the RF classifier trained and optimised on the *Arabidopsis* data.

Identification of SCW-associated transcription factors and structural genes in *E. grandis*

Arabidopsis thaliana TFs associated with SCW biosynthesis were compiled from a list of 600 literature-supported interactions with SCW structural genes or transcription factors (Supplementary Dataset D3). This was originally derived from Data sheet 1 in Hussey et al. (2013) and updated. A non-redundant list of the implicated TFs was used to identify 86 one-to-one *E. grandis* orthologs via the integrative orthology tools in PLAZA 3.0 (Proost et al. 2015), first by assigning them to OrthoMCL gene families, then subjecting them to tree-based orthology, and finally to best BLAST hit analysis. This list was supplemented with TFs (that is, those annotated with GO term 0003700) associated with SCW processes in *Eucalyptus*, namely 16 TFs that were differentially expressed during tension wood formation in *E. grandis* × *E. urophylla* (Mizrachi et al. 2015) and 45 TFs occurring in SCW-enriched co-expression cluster PC3 from Pinard et al. (2019) (Supplementary Dataset D4). Structural genes involved in cellulose, hemicellulose and lignin biosynthesis in *E. grandis* were obtained from Myburg et al. (2014) and Carocha et al. (2015). These SCW-associated transcription factors and structural genes were used to annotate the target genes inferred using DAP-seq-ML (See full set of target genes and peaks in Supplementary Dataset D5).

In order to investigate the conservation of *Eucalyptus* MYB downstream targets, we compared *Egr*MYB2 and *Egr*MYB137 target genes identified by DAPseq-ML with the set of genes targeted by their closest orthologs: *At*MYB83 in *Arabidopsis* and *Ptr*MYB074 in *Populus*. We used OrthoVenn3 (at https://orthovenn3.bioinfotoolkits.net/) to construct robust genes orthogroups from the genomes of *Arabidopsis thaliana, Populus trichocarpa, Eucalyptus grandis, Medicago truncatula, Cucumis sativa, Beta vulgaris, Ipomoea triloba and Prunus persica* (Phytozome V13). Each orthogroup enclosed several *Arabidopsis, Populus* and *Eucalyptus* genes considered to be the closest orthologs. Orthogroups were used to compare *Egr*MYB2 and *Egr*MYB137 targets in *Eucalyptus*, with PtrMYB074's targets in *Populus* and AtMYB83 in *Arabidopsis*. The comparison of *Egr*MYB137 DAP-seq-ML targets with target genes of *Ptr*MYB074 identified by ChIP-seq (Liu et al. 2022) and *Egr*MYB2 DAP-seq-ML targets with *At*MYB83 ampDAP-seq target genes (O'Malley et al. 2016) is available in Supplementary Datasets D5.

Supplementary Methods S3

FT-IR analysis

Eucalyptus hairy root samples and stems from transgenic poplars were grinded to powder, freeze-dried and extractives were removed by successive baths of boiling water, 100% ethanol, toluene/ethanol (50/50, v/v) and acetone. Extractive-free xylem residues were analysed by Fourier Transformed infra-red spectroscopy using an attenuated total reflection (ATR) Nicolet 6700 FT-IR spectrometer (Thermo Fisher) equipped with a deuterated-triglycine sulfate (DTGS) detector. The analysis was led on 20 p35S:EgrMYB137 Eucalyptus hairy roots independent lines and 8 independent lines of empty vectors; and 15 p35S:EgrMYB137-EAR poplar samples (3 independent lines x 5 biological replicates) and 6 empty vectors (3 independent lines x 2 biological replicates). Spectra were acquired between 4000 and 400 cm-1 with a 4 cm-1 resolution and 32 scans per spectrum. For each sample, ten measures were performed and a resulting median spectrum of the ten technical replicates was calculated. Baseline correction, normalization and offset correction was performed using R packages (hyperspectr, prospect and base respectively). Partial Least Square-Discriminant analysis (PLS-DA) was performed using mixOmics R package (Rohart et al. 2017) to compare samples. To allow a unique comparison between poplar and Eucalyptus samples using 3D-PLS-DA, we subtracted the median of the empty vectors samples to each individual spectrum, either EgrMYB137 samples or empty vectors. Sparse-PLSDA was used to identify the 200 most discriminant wavenumbers on PC1 and PC2.

References

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

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

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

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186. https://doi.org/10.1093/nar/gkr944

Hussey SG, Mizrachi E, Creux NM, Myburg A a (2013) Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. Front Plant Sci 4:325. https://doi.org/10.3389/fpls.2013.00325

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

Mizrachi E, Maloney VJ, Silberbauer J, Hefer CA, Berger DK, Mansfield SD, Myburg AA (2015) Investigating the molecular underpinnings underlying morphology and changes in carbon partitioning during tension wood formation in Eucalyptus. New Phytologist 206:1351–1363. https://doi.org/10.1111/nph.13152

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

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

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295. https://doi.org/10.1038/nbt.3122

Pinard D, Fierro AC, Marchal K, Myburg AA, Mizrachi E (2019) Organellar carbon metabolism is coordinated with distinct developmental phases of secondary xylem. New Phytologist 222:1832–1845. https://doi.org/10.1111/nph.15739

Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inzé D, Mueller-Roeber B, Vandepoele K (2015) PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res 43:D974–D981. https://doi.org/10.1093/nar/gku986

Rohart F, Gautier B, Singh A, Lê Cao K-A (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 13:e1005752. https://doi.org/10.1371/journal.pcbi.1005752

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

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

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