Identification and functional evaluation of accessible chromatin associated with wood formation in *Eucalyptus grandis*

Katrien Brown[1], Lazarus T. Takawira[1], Marja M. O'Neill[1], Eshchar Mizrachi[1], Alexander A. Myburg[1] and Steven G. Hussey[1*]

[1]Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), Genomics Research Institute (GRI), University of Pretoria, Private Bag X28, Pretoria, 0002, South Africa

*Corresponding author: +2712 420 3294 / steven.hussey@fabi.up.ac.za

Word count (body): 6,191
Word count (summary): 199
Word count (introduction): 1,001
Word count (materials and methods): 1,734
Word count (results): 2,076
Word count (discussion): 1,276
Word count (acknowledgements): 104
Number of figures: 7 (all in colour)
Number of tables: 1 (one in colour)
Number of supporting information documents: 1

Summary

- Accessible chromatin changes dynamically during development and harbours functional regulatory regions which are poorly understood in the context of wood development. We explored the importance of accessible chromatin in *Eucalyptus grandis* in immature xylem generally, and MYB transcription factor-mediated transcriptional programmes specifically.
- We identified biologically reproducible DNase I Hypersensitive Site (DHSs) and assessed their functional significance in immature xylem through their associations with gene expression, epigenomic data and DNA sequence conservation. We identified *in vitro* DNA binding sites for six secondary cell wall-associated *Eucalyptus* MYB (EgrMYB) transcription factors using DAP-seq, reconstructed protein-DNA networks of predicted targets based on binding sites within or outside DHSs and assessed biological enrichment of these networks with published datasets.
- 25,319 identified immature xylem DHSs were associated with increased transcription and significantly enriched for various epigenetic signatures (H3K4me3, H3K27me3, RNA pol II), conserved noncoding sequences and depleted single nucleotide variants. Predicted networks built from EgrMYB binding sites located in accessible chromatin were significantly enriched for systems biology datasets relevant to wood formation, while those occurring in inaccessible chromatin were not.
- Our study demonstrates that DHSs in *E. grandis* immature xylem, most of which are intergenic, are of functional significance to gene regulation in this tissue.

Keywords: accessible chromatin, DNase I hypersensitivity, immature xylem, DNase-seq, *Eucalyptus*, wood formation

Introduction

The analysis of the genome of the widely grown lignocellulosic feedstock, *Eucalyptus grandis* (Myburg *et al.*, 2014) and accompanying transcriptomics, marker mapping and population genetics studies has equipped forest scientists with a wealth of genomic tools to understand and improve woody biomass production in hardwood trees. While functional genomics studies and rapidly accumulating transcriptomic data continue to assist with the discovery and functional annotation of wood formation genes in *E. grandis*, a remaining challenge is the identification and evaluation of functional noncoding sequences that regulate gene expression. DNA elements such as enhancers and insulators act independent of distance and orientation to their target genes and are particularly elusive (Narlikar & Ovcharenko, 2009). This so-called "dark matter" (Jiang, 2015) of plant genomes typically harbours, among others, *cis*-regulatory elements (CREs) that regulate transcription and rely on an open chromatin context appropriate to a given tissue (Song *et al.*, 2011). Chromatin-based signatures can thus be used to identify new promoters and enhancers with high sensitivity (Heintzman *et al.*, 2007; Zhu *et al.*, 2015; Daugherty *et al.*, 2017; Lu *et al.*, 2018). Since the discovery that nuclease cleavage is affected by protein binding (Galas & Schmitz, 1978), accessible chromatin regions that are hypersensitive to DNase I cleavage (DNase I hypersensitive sites; DHSs) have also been found to accurately indicate protein binding and have a strong association with active promoters,

enhancers, insulators, suppressors and specific histone modifications (Gross & Garrard, 1988; Boyle *et al.*, 2011; Thurman *et al.*, 2012; Calo & Wysocka, 2013; Weber *et al.*, 2016).

The ENCODE (Encyclopedia of DNA Elements) and model organism ENCODE (modENCODE) projects have identified active CREs in numerous tissues of humans and model organisms, respectively (Gerstein *et al.*, 2010; The modENCODE Consortium, 2010; The ENCODE Project Consortium, 2012). While no equivalent projects currently exist for plants, several genome-wide DNase-seq, DNA methylation and histone modification studies have been conducted in *Arabidopsis thaliana*, *Zea mays, Populus trichocarpa*, *Solanum lycopersicum* and *Oryza sativa*, for example (Zhang *et al.*, 2006; Zilberman *et al.*, 2007; Li *et al.*, 2008; Zhang *et al.*, 2009; Weinhofer *et al.*, 2010; Zhang, Wenli *et al.*, 2012; Du *et al.*, 2013; Gent *et al.*, 2014; Pajoro *et al.*, 2014; Sullivan *et al.*, 2014; Cumbie *et al.*, 2015; Qiu *et al.*, 2016; Rodgers-Melnick *et al.*, 2016; Engelhorn *et al.*, 2017; Oka *et al.*, 2017). In one interesting study, DNA methylation was analyzed specifically in DNase I accessible chromatin of poplar shoot apical meristems through tandem DNase I treatment followed by methylated cytosine ChIP-seq (Lafon-Placette *et al.*, 2013). However, few studies have examined the epigenomic landscape of vascular tissues (e.g. Hussey *et al.*, 2015; Wang *et al.*, 2016; Hussey *et al.*, 2017) and only one study has identified genome-wide accessible chromatin regions in the vascular cambium, in this case as part of an integrated co-expression network study of wood formation (Zinkgraf *et al.*, 2017). The latter revealed physical enrichment of DHSs for genes in co-expression modules that are conserved across diverse

environments and linked to SNPs associated with wood chemistry and biomass-related traits. While this result shows that DHSs may be important in gene co-expression networks underlying wood formation, the study did not attempt to demonstrate biological replication of the DHSs, assess selective constraints at these regions or uncover possible mechanisms underlying the link between DHSs and the regulation of transcriptional modules linked to wood quality traits.

Developmental programmes related to wood formation, such as secondary cell wall (SCW) formation, lignification and programmed cell death, are regulated by at least three tiers of transcription factors (TFs) involving principally MYB- and NAC-domain regulators (reviewed by Hussey *et al.*, 2013; Li *et al.*, 2016; Zhang *et al.*, 2018). In *Arabidopsis*, the NAC master regulators VND6, VND7, NST1, NST2 and NST3 activate second-level TFs MYB46 and MYB83, among others, which act as hub nodes for the activation of other MYB TFs (e.g. MYB20, MYB58, MYB63, MYB85), NAC TFs (e.g. KNAT7) and SCW structural genes (Mitsuda *et al.*, 2007; Zhong *et al.*, 2008; McCarthy *et al.*, 2009; Zhou *et al.*, 2009; Li *et al.*, 2011; Bhargava *et al.*, 2013; Ko *et al.*, 2014). While there is some evidence that several TFs within the *Arabidopsis* network are functionally conserved in woody plants (Zhong *et al.*, 2010; Zhong *et al.*, 2011; Kalluri *et al.*, 2014; Nakano *et al.*, 2015), the presence of woody plant-specific subgroups of MYB TFs with proven roles in regulating xylem development (Soler *et al.*, 2015; Soler *et al.*, 2016) and a large number of novel protein-DNA interactions in model woody plants (Chen *et al.*, 2019) suggests that SCW transcriptional networks undergo rewiring in different lineages. Systems genetics and co-expression network analyses in *Eucalyptus*, which is emerging as

an important model for wood formation, are implicating new genes in certain aspects of wood formation or linking them to lignocellulosic biomass traits (Myburg *et al.*, 2014; Dharanishanthi & Dasgupta, 2016; Mizrachi *et al.*, 2017; Pinard *et al.*, 2019; Ployet *et al.*, 2019). However, the mechanisms by which they are transcriptionally regulated through interactions with CREs in the context of accessible chromatin in developing xylem is not yet understood.

In this study, we employed DNase-seq to comprehensively explore the chromatin landscape of *E. grandis* developing (immature) secondary xylem. We critically evaluated different DHS detection algorithms, employed stringent biological reproducibility measures and assessed DHSs identified from differing DNase-seq fragment length distributions. We provide evidence of functional enrichment from gene expression, modified histone, conserved noncoding sequence (CNS) and single nucleotide variant (SNV) data, and demonstrate that immature xylem DHSs, being associated with ~26% of annotated genes, can be used to significantly enrich SCW-related EgrMYB-target gene networks for network modules linked to wood quality traits. The immature xylem DHSs identified in this study also serve as an important resource for the functional annotation of the *E. grandis* noncoding genome, the prioritization of functionally relevant single nucleotide polymorphism (SNP) markers, the computational prediction of transcriptional networks and understanding the multi-dimensional control of gene expression during xylogenesis (Ritchie *et al.*, 2014; Van de Velde *et al.*, 2014; Banf & Rhee, 2017).

Materials and methods

Tissue collection

For DNase-seq, we used immature secondary xylem samples previously collected from three eight-year-old ramets of *Eucalyptus grandis* clone TAG0014 in Kwambonambi, South Africa (Hussey *et al.*, 2017). For whole-genome re-sequencing, young leaf and mature xylem were collected from different TAG0014 ramets. To verify their clonal identities, the trees were genotyped using a set of ten custom microsatellite markers (Byrne *et al.*, 1996; Brondani *et al.*, 1998; Brondani *et al.*, 2002) and compared to a pre-established consensus clonal DNA fingerprint for TAG0014.

Nuclear isolation

Plant tissue was ground to a fine powder with liquid nitrogen using an A 11 analytical grinder (IKA Works GmbH & Co., Staufen im Breisgau, Germany) followed by fine grinding with a mortar and pestle. Nuclei were isolated according a protocol adapted from Cumbie *et al.* (2015). Briefly, the ground tissue (~9 g) was added to 100 ml of Nuclear Isolation Buffer (NIB; 20 mM Tris-HCl, 25 mM EDTA, 40% (w/v) glycerol pH 7.5, 10 mM spermidine, 0.1% 2-mercaptoethanol) by gently stirring the mixture on ice for 5 minutes. The solution was successively filtered through a 100 µm and 60 µm nylon filter mesh. The filtrate was divided into 50 ml falcon tubes before being centrifuged for 10 minutes at 1800 × *g* at 2 °C. The pellets were gently resuspended in 15 ml NIB supplemented with 0.5% Triton X-100 and complete Protease Inhibitor Cocktail (Roche) and centrifuged at

1800 × *g* for 8 minutes (repeated twice). The resultant nuclear pellet was rinsed by resuspension in 20 ml RSB buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$) followed by centrifugation at 1100 × *g* for 10 minutes. Nuclei were digested, DNA extracted and libraries prepared as described in Method S1. Briefly, nuclear extracts were subdivided and treated with four DNase I concentrations (0U, 15U, 25U and 40U) to allow for selection of optimally sheared chromatin. The reactions were immediately flash-frozen (1 min) to permeate the nuclear membrane, and allowed to thaw for 15 min at 37 °C. The nuclease digestions were attenuated with EDTA, and genomic DNA was extracted from the pellets and size-selected on an agarose gel (50 – 150 bp for small-fragment libraries; 150 – 300 bp for large-fragment libraries) across all DNase I concentrations to account for batch-to-batch variation in DNase I cleavage efficiency. DNase-seq libraries were constructed by Novogene, Inc. USA) and SE50 sequencing performed on the HiSeq2500 platform.

Whole-genome re-sequencing and construction of a custom *E. grandis* sequence

Genomic DNA was extracted from young leaf and xylem samples (Qiagen DNeasy Plant Kit) and sequenced (PE100) on a HiSeq2500 (Beijing Genome Institute, Hong Kong) to >30X coverage of the *E. grandis* genome (Myburg *et al.*, 2014). Read mapping and variant detection was performed in Galaxy (Goecks *et al.*, 2010) with input parameters given in Table S1. First, 66,064,029 paired-end xylem sequences and 68,833,129 paired-end young leaf sequences were merged and mapped to the *E. grandis* v.1.1. reference genome (Myburg *et al.*, 2014) using the Burrows-

Wheeler Aligner v0.6.1 (BWA, Li and Durbin 2009). Paired Read Mate Fixer (Galaxy v1.56.0, Picard Tools) was used to ensure that mate-pair information was coordinated for paired reads, followed by re-alignment of reads around small Insertions and Deletions (InDels) using RealignerTargetCreator (Galaxy v0.0.4, GATK) and IndelRealigner (Galaxy v0.0.6, GATK) and flagging of PCR duplicates with MarkDuplicates (Galaxy v1.56.0, Picard Tools). MPileup (Galaxy v0.0.1, SAMtools) was used to obtain a mapped read summary and generate a variant file, while coverage was determined using BEDtools coverageBed (Galaxy v0.0.1, BEDtools). The MPileup output BCF file was converted to VCF format and high-confidence SNVs and small InDels were filtered using BCFtools (Galaxy v0.0.1, SAMtools).

The predicted genotypic ratios from the VCF file were used to determine whether a SNV differing from the BRASUZ1 reference (Myburg *et al.*, 2014) should be substituted. If the reference allele was present in the called TAG0014 variant, the reference variant was retained. If the most likely TAG0014 genotype was homozygous for an alternative allele to the reference, then the reference was substituted with the TAG0014 allele. For positions where two alternative alleles differing from the reference allele were present in TAG0014, the allele associated with the largest Alt_1Alt_1 and Alt_1Alt_2 predicted genotypic values (represented by a larger number of mapped reads) was substituted.

DNA affinity purification sequencing (DAP-seq)

Coding sequences of EgrMYB20 (Eucgr.F02864), EgrMYB46 (Eucgr.B03684), EgrMYB52 (Eucgr.F02756), EgrMYB83 (Eucgr.G03385), EgrMYB85 (Eucgr.D02014) and EgrMYB103 (Eucgr.D01819) from a synthetic gene panel (Hussey *et al.*, 2019) were cloned into pCR8©/GW/TOPO™ (Invitrogen) and then transferred to the pIX-HALO expression vector (The Arabidopsis Interactome Mapping Consortium, 2011). HaloTag-TF fusion proteins were expressed *in vitro* (TNT SP6 Coupled Wheat Germ Extract System, Promega), confirmed via Western Blot and quantified via semi-quantitative dot blot analysis. DAP-seq binding assays were performed as described by Bartlett *et al.* (2017) using 50 – 500 ng protein and 50 ng adaptor-ligated genomic DNA prepared from *E. grandis* clone TAG0014 immature xylem sonicated to ~200 bp with a Covaris S2 sonicator (Covaris Inc., Woburn, MA). A negative control library was prepared using HaloTag standard protein (Promega). Eluted DNA was PCR-amplified with Illumina TruSeq Index primers (21 cycles), pooled, size-selected (~200 to 400 bp) on an agarose gel, and sequenced on an Illumina MiSeq.

Bioinformatics and statistical analysis

The *E. grandis* v1.1 TAG0014 custom reference genome was indexed using BWA (Li & Durbin, 2010). Raw DNase-seq sample reads (50 nt) were assessed for quality using FastQC (Andrews, 2010) and aligned to the custom reference using BWA. Duplicates and reads of poor quality (MAPQ < 30) were flagged and filtered out using Samtools (Li *et al.*, 2009). To evaluate different peak-callers, peaks in

three samples IX1-small, IX2-small and IX3-small were identified with a lenient threshold using MACS2 (Zhang *et al.*, 2008; *P*-value < 0.05), Hotspot2 (Sabo *et al.*, 2004; FDR < 0.8), and F-seq (Boyle, Alan P. *et al.*, 2008; FDR < 0.05). The peak sets were then compared using the IDR method (see Note S1) to establish the biological reproducibility of peaks identified within and between peak callers. Once optimized, IX-DHS peaks were predicted at a lenient threshold (*P*-value =< 0.05) using MACS2 (Zhang *et al.*, 2008) with a 3' to 5' shift of 25 bp (i.e. the midpoint of a 50 bp read) in order to correct for the read coverage relative to the DNase I cut site. The TAG0014 re-sequencing data was used as the background control. Peaks predicted from the DNase I digested naked DNA control (NDC) were used as a reference dataset for potential DNase I cleavage bias; any chromatin DNase-seq peaks overlapping an NDC peak by at least 1 bp were discarded. After assessing biological reproducibility and determining an experiment-wide threshold (Note S1), DNase-seq datasets across biological replicates were merged into a single BAM file, the peaks were called again using MACS2 (*P*-value < 0.05) and then ranked by *P*-value followed by application of the experiment-wide cut-off threshold to arrive at a final set of 25,319 peaks.

The *E. grandis* v1.1 annotation was used as reference for the genomic location of gene features (5' UTR, exons, introns and 3' UTR) and 1.5 kb promoter regions. The distribution of peak summits across genomic features was determined using BEDtools intersectBed (Quinlan & Hall, 2010). Genes overlapping DHSs were defined based on any overlap with the gene annotation (transcribed region) or 1.5 kb upstream of the transcribed region. Corresponding RNA-seq data was derived

from For SNV density analysis, DHSs were centered around their summits and artificially extended 2500 bp to either side using BEDtools, discarding overlapping extended DHSs. SNV count at each position of extended DHSs was determined using BEDtools and normalized according to number of DHSs.

Histone modification enrichment was assessed by the percentage of DHSs overlapping the histone marks (from Hussey *et al.*, 2017), determined by BEDtools intersectBed. An analysis of IX- and NDC DHS proximity to H3K4me3 and H3K27me3 histone modification sites was performed using the summit of each histone modification site as the centerpoint, determining the number of DHS summits falling within each 50 bp sliding window from the summit and normalizing the results for the number of histone modification sites. A random sample was produced by BEDtools shuffleBed on the IX-DHS datasets and mapped as before, taking the median of 1,000 iterations. The number of CNSs overlapping DHSs was determined using BEDtools intersectBed and, using coverageBed, the coverage over 2,500 bp to either side of DHS summits by CNSs was determined.

DAP-seq analysis, MiSeq reads were mapped to the custom *E. grandis* TAG0014 genome described above using Bowtie2 (Langmead & Salzberg, 2012). Duplicated reads were discarded. Significant binding event and motif discovery was conducted using GEM v2.5 (Guo *et al.*, 2012) with parameters --f SAM --k_min 6 --kmax 20 --k_seqs 600 --k_neg_dinu_shuffle. Binding sites were assessed for IX-DHS and NDC DHS enrichment relative to the median of 1000 randomly shuffled binding sites for each TF dataset. A TF-target gene network of the six EgrMYB TFs

was constructed by assigning target genes to binding sites if they fell within the 1.5 kb promoter or transcribed region, the latter considered possible given that TF binding to exons is widespread (Sullivan *et al.*, 2015). In ambiguous cases where more than one gene met these criteria (~12% of cases), both genes were regarded as potential targets. A power-law was fitted to the node degree distribution using fit_power_law() (https://igraph.org/). Genes associated with DHSs that co-located with the HaloTag DAP-seq library (n = 122) served as a negative control. Hypergeometric tests of genes in each network were conducted with respect to xylogenesis-related dataset enrichment. Four *E. grandis* xylogenesis-related datasets were considered, namely 1,597 genes associated with thirteen complex wood property traits from a network-based systems genetics study (Dataset S1 from Mizrachi *et al.*, 2017), a curated list of 542 *E. grandis* homologs of SCW-related reference genes (Dataset S2 from Mizrachi *et al.*, 2017), 1,047 genes across three co-expression modules (PC3, MC3, MC5) from a xylem organellar carbon allocation study with significant enrichment for SCW-related terms (Pinard *et al.*, 2019), and 377 genes across six co-expression modules (M1-M6) from an abiotic stress-related xylem systems biology study that were enriched for xylogenesis-related terms and significantly correlated with woody traits such as vessel density, vessel diameter and saccharification yield (Ployet *et al.*, 2019). A set of 363 genes across six non-xylogenesis-related co-expression modules for the latter study (M7-M12) was used as a negative control for specific enrichment for xylogenesis.

Results

Identification of DNase I hypersensitive sites in immature xylem

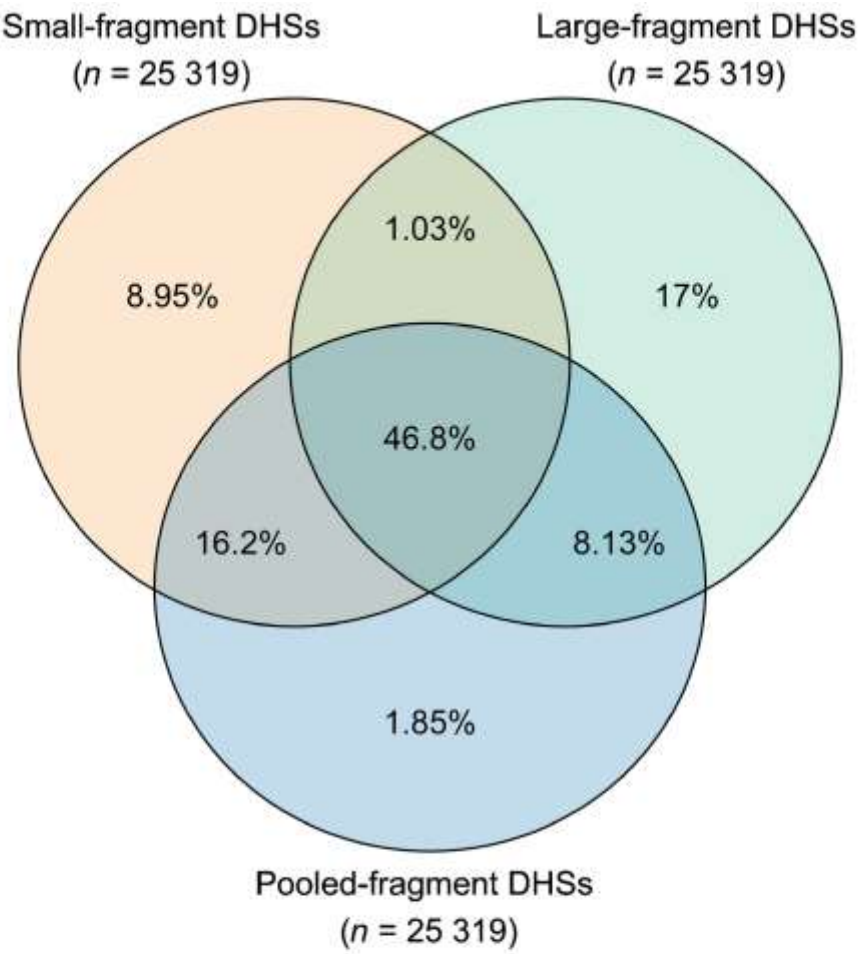
We implemented a DNase-seq protocol to detect genome-wide accessible chromatin in developing xylem tissue from three ramets (IX1, IX2, and IX3) of an *E. grandis* genotype (clone TAG0014) grown in the field. In optimisation experiments, DNA extracted from untreated chromatin was of high molecular weight, while chromatin digested with various amounts of DNase I (15U, 25U and 40U) yielded desired fragment smears ranging from high (several kb) to low (~50 bp) molecular weight (Fig. S1a).

He *et al.* (2014) and Vierstra *et al.* (2014) reported that small digested fragments (<125 bp) are more likely to originate from DHSs in nucleosome-free regions where TF binding sites have been exposed while larger fragments (126–185 bp) spanning a ~147 bp nucleosome complex often originate from loosely packed chromatin rather than true nucleosome-free regions. To study these individually, we prepared two separate libraries for each sample consisting of DNase I-digested fragments of 50 bp to <150 bp (small-fragment libraries) and >150 bp to 300 bp (large-fragment libraries) (Fig. S1a). For large- and small-fragment libraries, DNase-seq libraries were prepared from digested chromatin from the three optimal DNase I concentrations (15U, 25U and 40U) and sequenced on an Illumina HiSeq 2500. As a control, genomic DNA extracted from IX1 was hydrolysed with low concentrations of DNase I (Fig. S1b) to account for potential DNase I cleavage biases in different sequence and DNA methylation contexts (Lazarovici *et al.*, 2013; Koohy *et al.*, 2014; Meyer & Liu, 2014).

The DNase-seq libraries yielded between ~36 and ~49 million reads (Table S2). After mapping the reads, we identified high-confidence, biologically meaningful DHSs by assessing the performance of three peak-calling software, MACS2 (Zhang *et al.*, 2008), Hotspot2 (Sabo *et al.*, 2004), and F-seq (Boyle, Alan P. *et al.*, 2008) (Table S3). The greatest mean number of peaks identified with a lenient threshold, by a large margin, was called by F-seq (Fig. S3a). Despite this, F-seq called the fewest biologically reproducible peaks (Fig. S3b). Hotspot2 and MACS2 identified a similar number of reproducible peaks but, to achieve this, a lenient false discovery rate threshold of 0.8 had to be applied initially in the case of Hotspot2 (Fig. S3a). Overall, the reproducibility of MACS2 peaks across biological replicates and across algorithms was clearly superior to Hotspot2 and F-seq (Fig. S3b). Furthermore, unlike F-seq, MACS2 can accommodate a control dataset (whole-genome re-sequencing data in this case) and was our preferred method for identifying DHSs.

The largest number of peaks that were reproducibly called at an IDR < 0.05 was 25,319 (see Note S1). Three bulked libraries were then compiled by combining mapped reads from (1) the small-fragment library replicates, (2) the large-fragment library replicates and (3) pooled-fragment fragment reads libraries, which we denoted as immature xylem DHSs (IX-DHSs). Based on physical overlap, 46.8% of the DHSs were shared between all three fragment size libraries, with the remainder of DHSs primarily originating from either the small- or large-fragment libraries (Fig. 1). The IX-DHS library contributed less than 2% of unique peaks,

(a)



(b)

Compared DHS sets	Jaccard index
Small-fragment vs Large-fragment	0.49
Small –fragment vs Immature xylem	0.76
Large-fragment vs Immature xylem	0.60

Figure 1. Overlap of *Eucalyptus grandis* DNase I hypersensitive site (DHS) datasets from different DNase-seq fragment size ranges. DNase-seq libraries consisting of small-fragments (50 to < 150 bp), large-fragments (> 150 to 300 bp) or pooled immature xylem fragments (50–300 bp) sizes were used to call three sets of DHSs; the DHS overlap percentage is indicated as a Venn diagram (a) and a Jaccard Index (b) of shared DHSs.

while the large-fragment library had the largest proportion of unique peaks (~17%). These results suggest that large-fragment and small-fragment DHSs represent biologically reproducible but contrasting chromatin biology, while the pooled-fragment (IX-DHS) library captured approximately 72% of the DHSs identified across small-fragment and large-fragment chromatin contexts.

Immature xylem DNase I hypersensitive sites occur preferentially near genes and transcription start sites

Because of their association with TF binding, DHSs tend to display a non-random distribution throughout the genome with a strong bias towards gene features, especially the TSS (Boyle, A. P. *et al.*, 2008; Zhang, Wenli *et al.*, 2012). To confirm whether this is true of IX-DHSs, we determined the distribution of the IX-DHS summits across genomic annotations. While most IX-DHSs were intergenic, the IX-DHSs were significantly enriched ($P < 0.05$, Fisher's exact test) for gene features, particularly promoters (1.2-fold), 5' UTRs (6.2-fold) and exons (2.7-fold), while being depleted in 3' UTRs (0.4-fold) and intergenic sequences (0.9-fold) (Fig. 2a). In contrast, the NDC DHSs were not significantly biased to any genomic features. The distribution of DHSs among genic features was further investigated with respect to DHSs unique to small- and large-fragment datasets (Fig. 2b). Unique small-fragment DHSs showed a significantly higher enriched in the 5'-UTRs (8.3-fold) and exons (2.6-fold), compared to unique large-fragment DHSs (2.3-fold in 5' UTRs, 1.7-fold in exons). Since the IX-DHS dataset was representative of the DHSs across small- and large-fragment chromatin contexts (Fig. 1), we concentrated further analyses on this dataset.

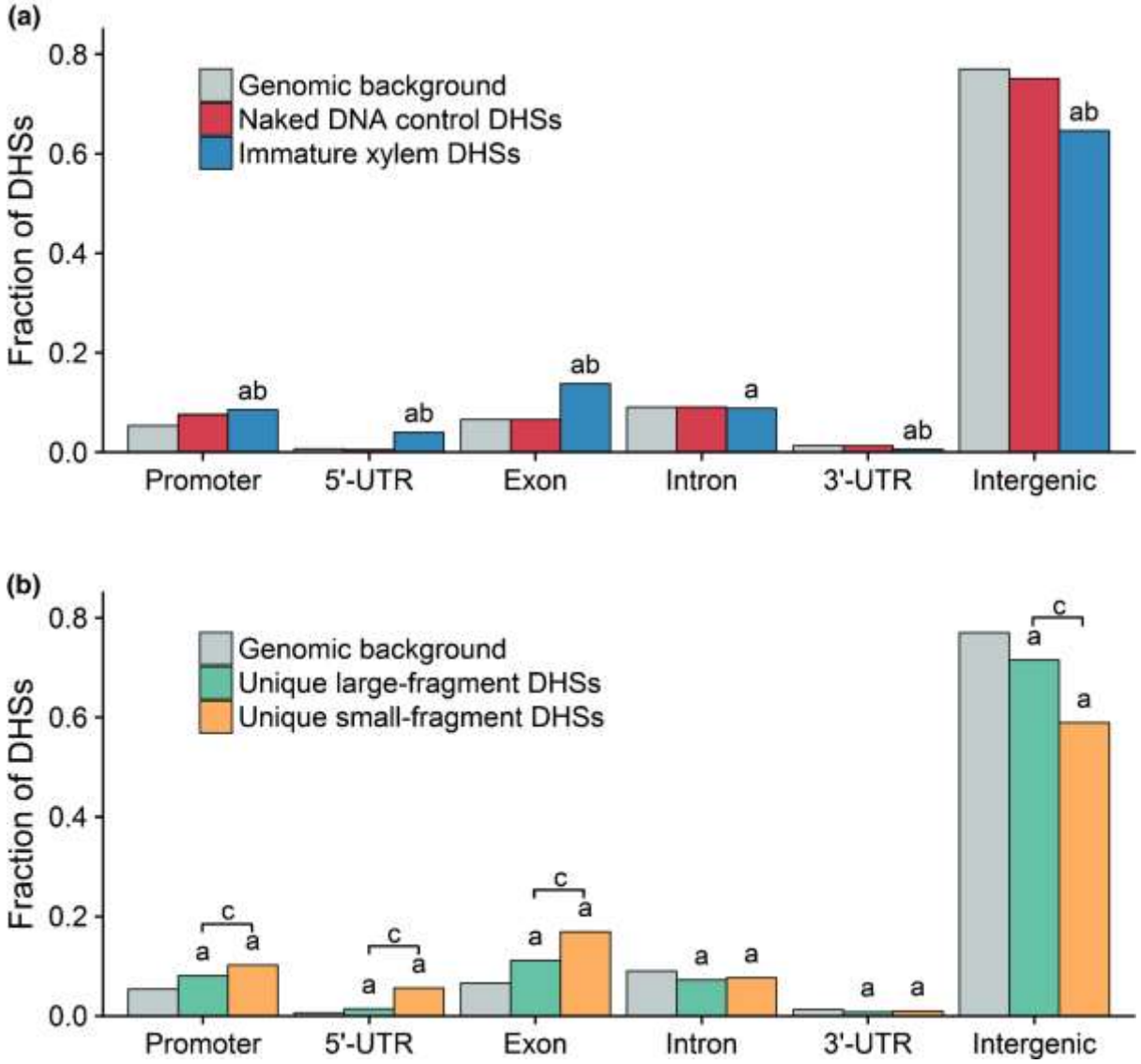


Figure 2. Distribution of DNase I hypersensitive sites (DHSs) across genomic annotations in *Eucalyptus grandis*. (a) Distribution of immature xylem DHSs and naked DNA control (NDC) DHSs (n = 25 319) across gene features. (b) Distribution of DHSs unique to small-fragment (n = 8731) and large-fragment DHS sets (n = 8651). The genomic background shows the proportion of each annotation relative to the genome size. Significance (Fisher's exact test with Bonferroni correction) indicated as [a]P-value < 0.05 relative to the genomic background, [b]P-value < 0.05 relative to NDC DHSs and [c]P-value < 0.05 between small- and large-fragment DHS datasets.

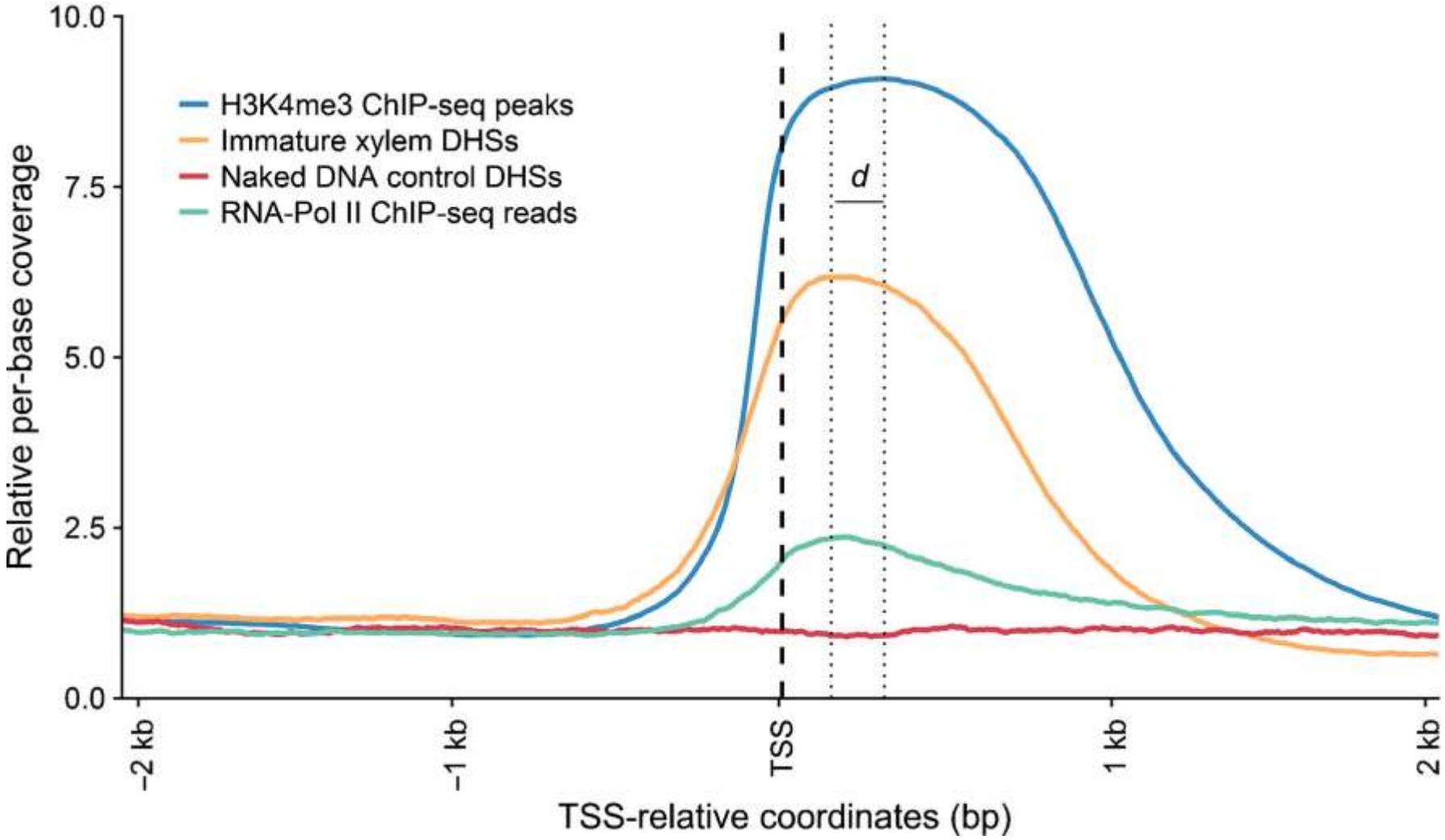


Figure 3. Per-base coverage of DNase I hypersensitive sites (DHSs) and modified histone peaks around predicted *Eucalyptus grandis* transcription start sites (TSSs). The normalized coverage of each base by modified histone ChIP-seq peak sets (H3K4me3, H3K27me3) and DNase-seq datasets (immature xylem, naked DNA control) was plotted relative to TSSs (n = 35 603), excluding alternate splice sites. The distance (d) between the coverage peaks of the IX-DHSs and H3K4me3 histone marks is *c*. 160 bp.

The association of IX-DHSs with proximal promoter regions was further investigated by comparing the coverage patterns of DHSs across transcription start sites (TSSs) with published RNA Pol II occupancy and H3K4me3 histone modification data (Hussey *et al.*, 2017) (Fig. 3). H3K4me3 and RNA Pol II occupancy peaked ~310 bp and ~190 bp downstream of the TSS, respectively. The IX-DHSs coverage summit occurred slightly upstream of RNA Pol II at ~170 bp downstream of the TSS, while NDC DHS coverage was relatively uniform 5,000 bp surrounding TSSs. An approximate distance of 160 bp separated the IX DHSs and H3K4me3 coverage summits (marked *d* in Fig. 3). While the sequence of peaks (DHSs preceding RNA Pol II, followed by H3K4me3) is in line with the role of accessible chromatin in the binding of TFs and transcriptional machinery to the promoter and TSS, it is surprising that the DHSs occurred downstream of the annotated TSSs. This result suggests that the TSS is misannotated in a large number of cases.

Enrichment of immature xylem DNase I hypersensitive sites for functional genomic and epigenomic data

The relationship between DHSs and gene expression was investigated by comparing the immature xylem transcript levels of genes overlapping IX-DHSs to those that did not, using RNA-seq data derived from the same samples (Hussey *et al.*, 2017). Compared to a negative control of genes overlapping NDC DHSs (n = 6,284; median FPKM 2.67) which produced a slight but significantly shifted distribution to those that did not (n = 30,093; median FPKM 0.79; *P*-value ≈ 0,

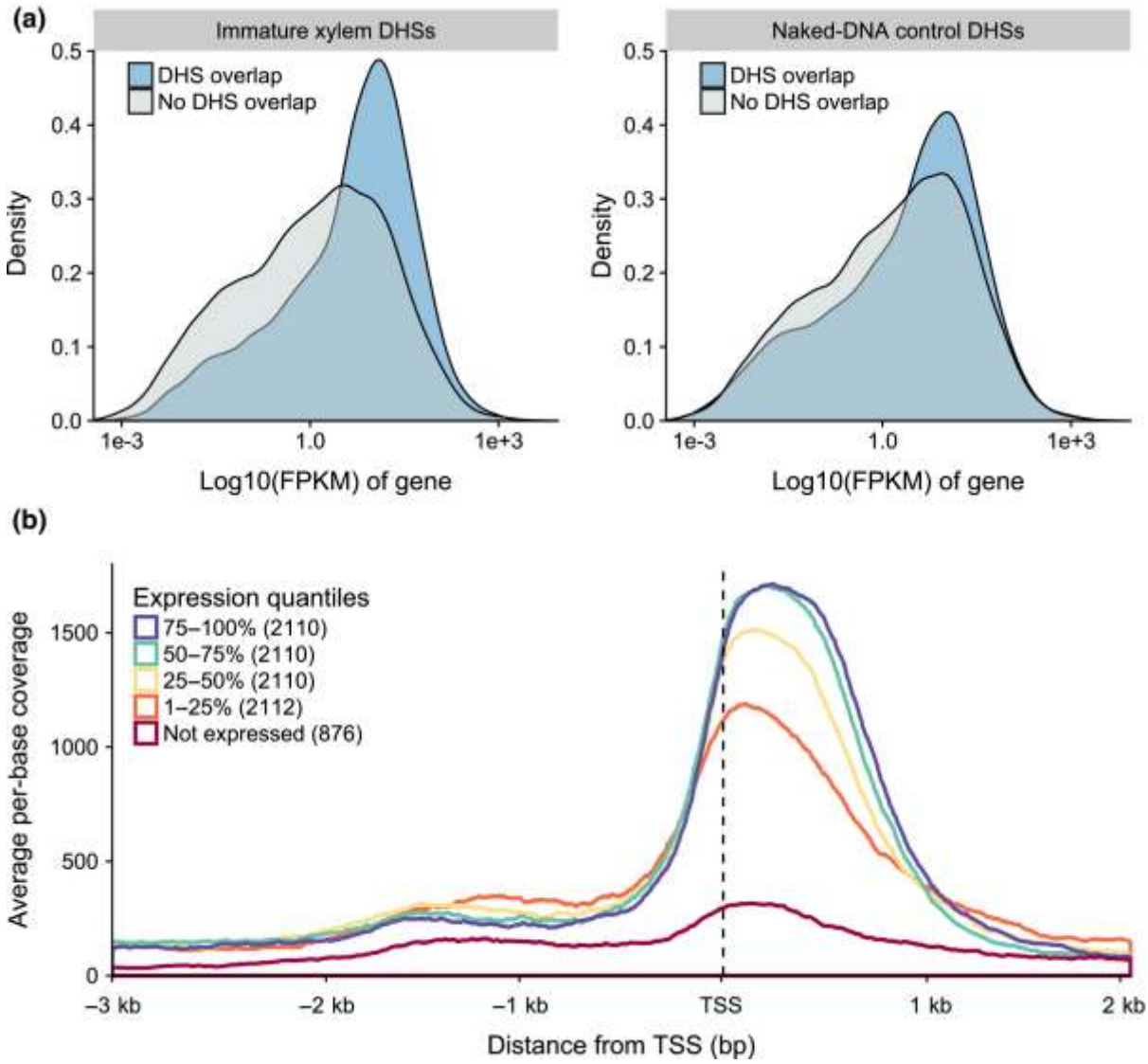


Figure 4. Immature xylem DNase I hypersensitive sites (DHSs) are positively associated with gene transcription in *Eucalyptus grandis*. (a) The kernel density estimation of $\log_{10}$(FPKM) values were compared for 9318 genes overlapping and 27 058 genes not overlapping immature xylem DHSs (left panel, *P*-value ≈ 0, Kolmogorov–Smirnov test). Distribution of genes overlapped (6283) and not overlapped (30 093) by naked DNA control DHSs (right panel, *P*-value ≈ 0). (b) Transcriptional start site (TSS) occupancy by DNase-seq reads for bins of genes overlapping DHSs divided within four quantiles of increasing transcript levels. FPKM, Fragments Per Kilobase of transcript per Million mapped reads.

Kolmogorov-Smirnov test), genes associated with IX-DHSs (n = 9,318; median FPKM 5.03) were expressed at far higher levels than those not overlapping DHSs (n = 27,058; median FPKM 0.52; *P*-value ≈ 0) (Fig. 4a). When we investigated the coverage of IX-DHSs near TSSs as a function of gene expression level in immature xylem, there was clearly a direct relationship between gene expression level and chromatin accessibility near the TSS (Fig. 4b), showing that the positioning of DHSs near the TSS has a particularly strong association with gene expression levels. The expression of IX-DHS-associated genes were similarly affected in six additional tissues as they were in immature xylem, however (Fig. S4). This suggests that most IX-DHSs are not exclusively associated with xylem development.

DHSs may co-occur with both activating (e.g. H3K4me3) and repressive (e.g. H3K27me3) histone marks (Shu *et al.*, 2011; Zhang, W. *et al.*, 2012). To test this for immature xylem DHSs, the association of DHSs with H3K4me3 and H3K27me3 ChIP-seq data derived from the same tissue samples (Hussey *et al.*, 2017) was investigated. Compared with NDC and randomly shuffled DHS data, significantly higher percentages of H3K4me3 (~23.6%) and H3K27me3 regions (~17.5%) overlapped IX-DHSs ($P << 10^{-100}$; Fisher's Exact Test) (Fig. 5a). Additionally, IX-DHS summits (i.e. the position of maximum local library coverage) preferentially occurred proximal to the summits of both H3K4me3 and H3K27me3 regions, while the NDC DHSs showed no such preference (Fig. 5b). The significant enrichment for overlap and strong spatial correlation between IX-DHSs and H3K4me3 regions is in line with the expectation that activating histone marks are

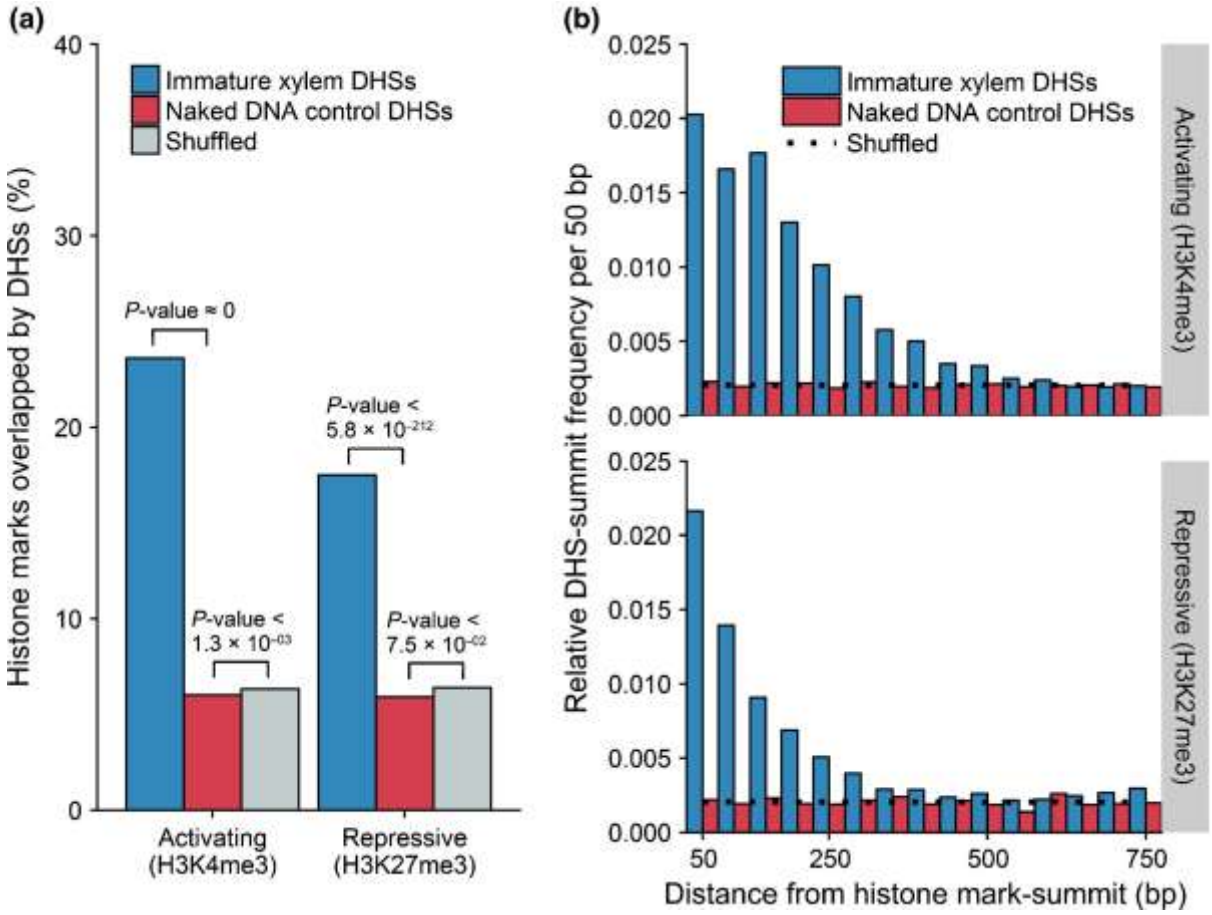


Figure 5. Association of DNase I hypersensitive sites (DHSs) with histone modifications in *Eucalyptus grandis* immature xylem. Physical overlap between genomic intervals occupied by the activating histone mark H3K4me3 (n = 31 784) or the repressive histone mark H3K27me3 (n = 16 690) and immature xylem DNase I-hypersensitive sites or naked-DNA control DHSs are shown. Shuffled DHSs represent the median outcome of 1000 permutations. (a) The percentage of DHS-overlapping histone modifications. (b) The relative frequency of DHS summits within 50-bp nonoverlapping windows of H3K4me3 summits (upper panel) and H3K27me3 summits (lower panel).

often associated with DHSs near genes to allow for protein binding. Furthermore, the small-fragment DHS dataset displayed a higher degree of proximal enrichment for H3K4me3 and H3K27me3 relative to the large-fragment dataset (Fig. S5), and are therefore likely more important in gene expression regulation.

CNSs are intimately associated with gene regulation, acting as conserved TF binding sites or other *cis*-acting regulatory elements (Hardison, 2000; Lockton & Gaut, 2005). It follows then that we expected immature xylem DHSs to be enriched for CNSs identified across ten dicot species including *E. grandis* (Van de Velde *et al.*, 2016), with the caveat that IX-DHS data is tissue-specific whereas the CNSs may function in many tissues and conditions. Some 4,567 of the IX-DHSs (18,0%) intersected with CNSs 12,933 times (a 4.5-fold enrichment; *P*-value ≈ 0, Fisher's Exact Test), compared to 1.3-fold enrichment for NDC DHSs (*P*-value = 1.8×10^{-10}) (Table S5). We also explored the proximal relationship between DHS summits and CNSs in the *E. grandis* genome. Relative to randomly shuffled and NDC DHS control datasets, DHSs from immature xylem showed a strong enrichment for CNSs proximal to the DHS summits across the genome (Fig. 6a). These results are consistent with the expectation that IX-DHSs are enriched for regulatory elements that participate in protein binding.

To explore local sequence evolution constraints on IX-DHSs, we re-sequenced a ramet of the *E. grandis* clone (TAG0014) sampled in this study and identified 1,519,673 single-nucleotide variants (SNVs) that differed from the BRASUZ1 reference genome as a proxy for naturally occurring SNP diversity. Interestingly,

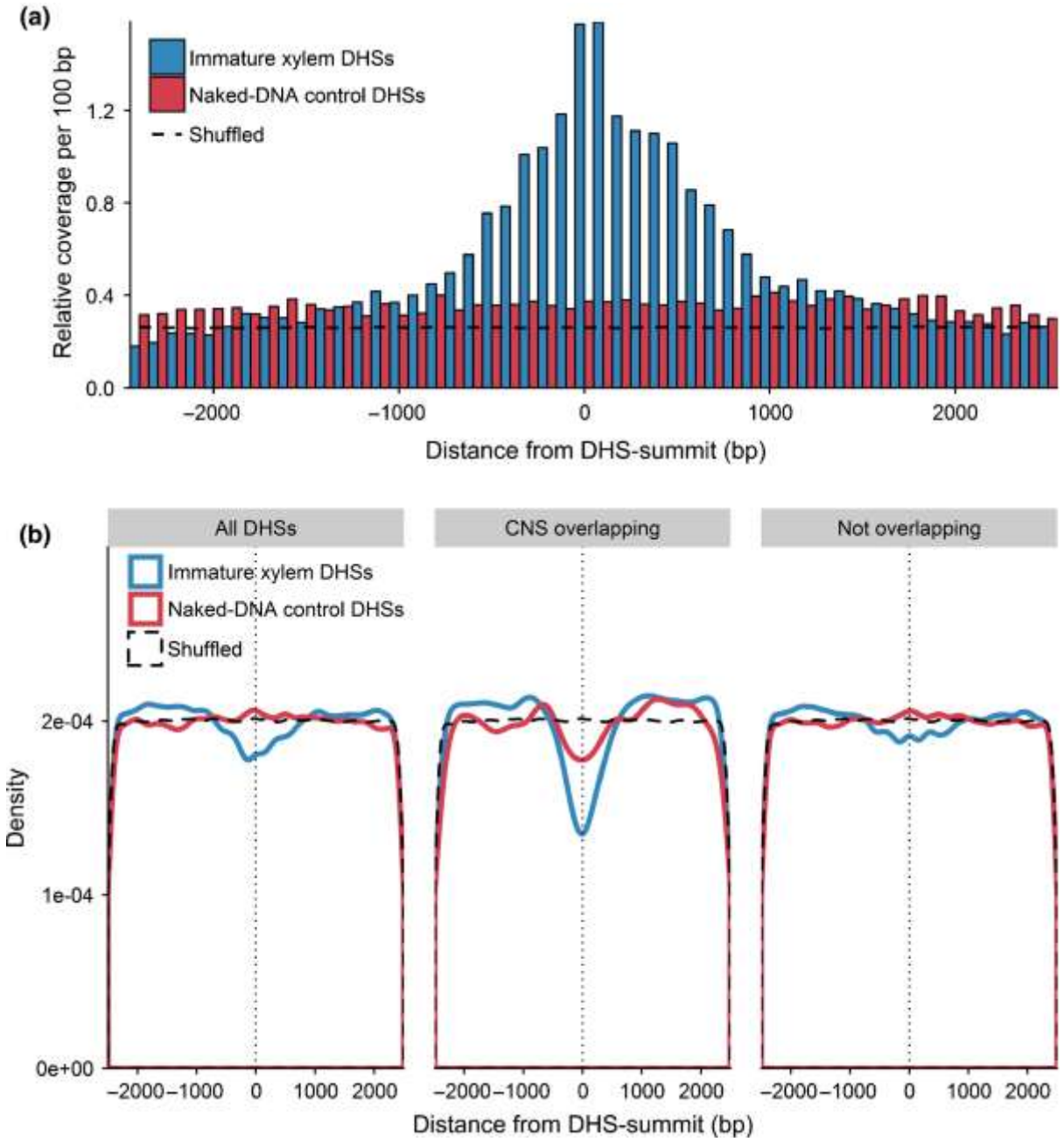


Figure 6. Association of DNase I hypersensitive sites (DHSs) with conserved noncoding sequences (CNSs) and single nucleotide variants (SNVs) in *Eucalyptus grandis* immature xylem. (a) The relative CNS coverage (per 100 bp bin) is plotted over a 5-kb window centred along the summits of the immature xylem DHSs and those of naked DNA control DHS and randomly shuffled DHS control datasets. (b) Relative SNV density centred on DHS summits in the *E. grandis* clone TAG0014 genome for all DHSs (left), DHSs overlapping transcriptional start site-proximal CNSs (centre; n = 4556 of immature xylem DHSs) and those that do not (right; n = 20 419 for immature xylem DHSs).

the IX-DHSs displayed a notable depletion of SNV density ±1000 bp surrounding the DHS summits, suggesting local purifying selection acting on the DHSs, whereas the NDC DHS and shuffled DHS control datasets revealed no relationship with SNV density as a function of proximity to their peaks (Fig. 6b, left). When DHSs were separately analyzed for SNV density depending on whether they overlapped TSS-proximal CNSs, CNS-overlapping IX-DHSs showed a more pronounced SNV depletion near their summits compared to those that did not overlap a CNS (Fig. 6b centre, right), further suggesting that an enrichment for functionally conserved CNSs may partially explain the SNV depletion observed within IX-DHS. Combined with the enrichment of IX-DHSs for other functional elements in the coding and noncoding genome (genes, gene expression, H3K4me3 regions, H3K27me3 regions and CNSs), this data supports the idea that the IX-DHSs identified in this study represent accessible chromatin of importance to gene regulation in *Eucalyptus*.

DNase I hypersensitive sites in immature xylem link woody biomass candidate genes to a regulatory network involving MYB transcription factors

Given the general importance of TSS-proximal DHSs in TF-mediated gene regulation, we aimed to better understand the relevance of accessible chromatin to TF-mediated regulation of genes linked to wood formation and lignocellulosic biomass traits. First, we identified genome-wide binding sites of six *E. grandis* homologs of SCW-associated MYB-domain TFs (Hussey *et al.*, 2013), named EgrMYB20, EgrMYB46, EgrMYB52, EgrMYB83, EgrMYB85 and EgrMYB103 after

their closest *Arabidopsis* homologs, using DAP-seq (Bartlett *et al.*, 2017). Between 3,536 and 16,414 significant binding events were identified for each EgrMYB TF (Table 1). Overrepresented motifs were identified within the EgrMYB TF binding sites for each candidate, most of which contained ACC-like sequences which are typical of MYB TFs (reviewed by Hussey *et al.*, 2013). Statistically, EgrMYB TF proteins bound within IX-DHS accessible regions more frequently than random chance, while the HaloTag negative control library was not associated with open chromatin (Table 1). This result is particularly interesting because DAP-seq assays identify *in vitro* binding events but nonetheless the EgrMYB TFs displayed a statistical preference for *in vitro* binding to sites within accessible chromatin in immature xylem.

Next, we reconstructed a TF-target gene network of the six EgrMYB TFs by considering binding events within accessible chromatin (IX-DHS regions) and assigning putative gene targets based on proximity of the binding site. This simple approach identified 899 potential target genes linked by five of the EgrMYB TFs. A second network was similarly reconstructed using EgrMYB binding sites that did not occur in accessible chromatin, yielding 6,347 putative target genes for the six EgrMYB TFs. The nodes of both networks displayed a power-law distribution typical of biological networks (Fig. S6). We then investigated the relevance of these networks to wood formation with respect to TF binding within IX-DHSs by calculating the enrichment of the genes in each network to various published SCW-related and non-SCW-related *Eucalyptus* datasets (see Methods). We found that the accessible chromatin-restricted network (i.e. EgrMYB-DHS-gene) was

Table 1. Enrichment of secondary cell wall-associated *E. grandis* MYB-domain transcription factors within immature xylem accessible chromatin. Significance was assessed using a Fisher's Exact Test of the observed DHS overlap relative to the median of 1000-fold randomly shuffled binding coordinates for each dataset.

	***Arabidopsis* ortholog**	**Number of significant binding events[a]**	**Top enriched motif**	**IX-DHSs[b]**		**NDC DHSs[c]**	
				Enrichment	*P*-value[d]	Enrichment	*P*-value
EgrMYB20	**MYB20**	9,424 (1,397)		5.92	$< 3.08 \times 10^{-16}$	1.11	0.410
EgrMYB85	**MYB85**	3,536 (434)		4.93	$< 3.08 \times 10^{-16}$	1.04	0.922
EgrMYB83	**MYB83**	16,414 (1,348)		3.27	$< 3.08 \times 10^{-16}$	1.01	0.927
EgrMYB46	**MYB46**	5,962 (424)		2.85	$< 3.08 \times 10^{-16}$	0.94	0.758
EgrMYB52	**MYB52**	7,386 (385)		2.07	$< 3.08 \times 10^{-16}$	1.15	0.325
EgrMYB103	**MYB103**	3,783 (135)		1.42	0.010	0.96	0.924
HaloTag (-)	-	13,039 (302)		1.18	0.054	1.04	0.775

[a]The numbers in parentheses indicate those located in accessible chromatin

[b]Immature xylem DNase I hypersensitive sites

[c]Naked DNS control DNase I hypersensitive sites

[d]Adjusted P-value (Benjamini & Hochberg, 1995)

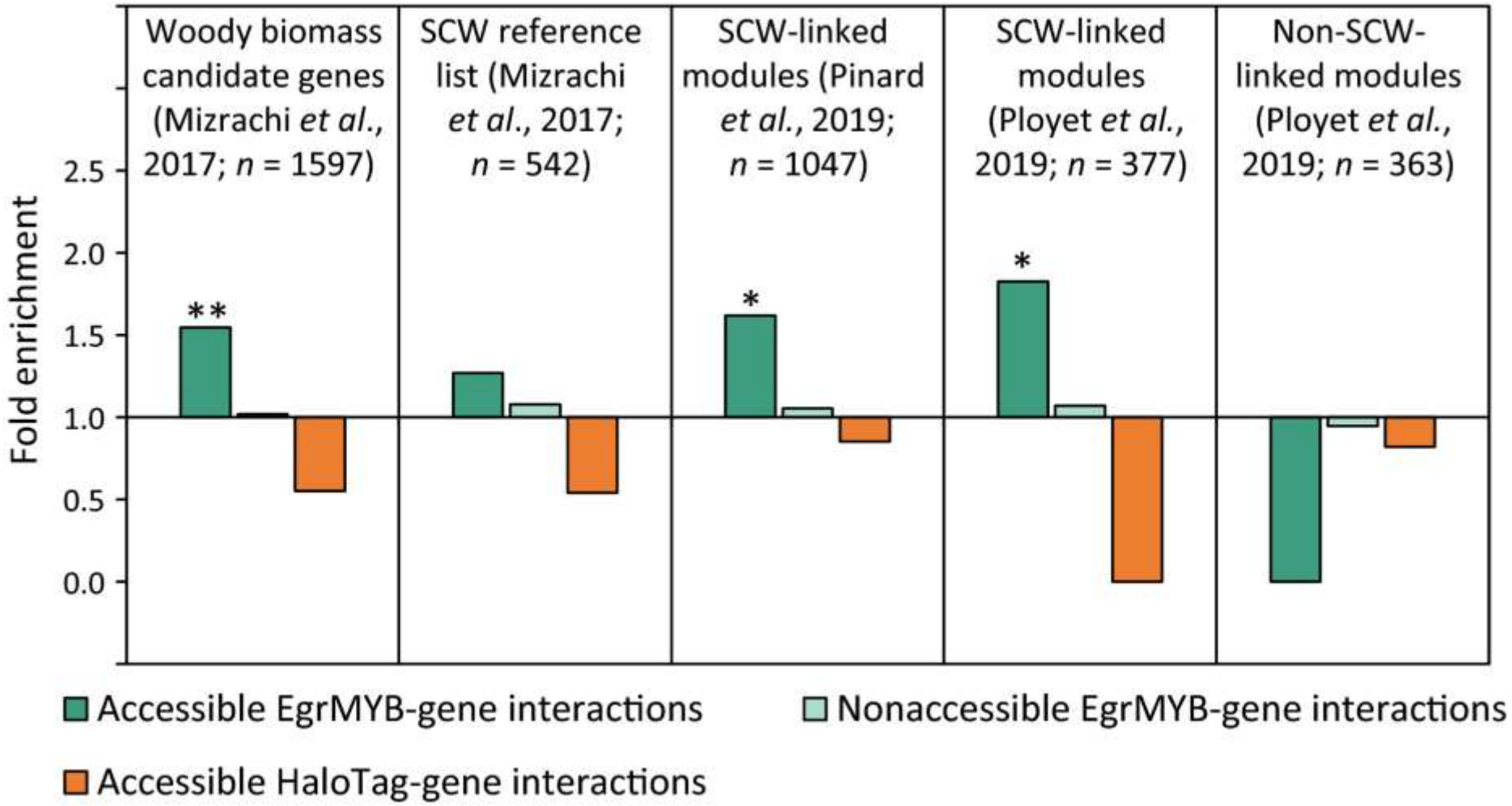


Figure 7. Enrichment of potential target genes of *Eucalyptus grandis* MYB (EgrMYB) transcription factors linked by DNase I hypersensitive sites (DHSs) in immature xylem for datasets associated with wood development and lignocellulosic biomass traits. Accessible EgrMYB-gene and HaloTag-gene interactions represent those occurring within IX-DHS regions. Adjusted *P*-values: *, $P < 0.05$; **, $P < 0.01$. SCW, secondary cell wall.

significantly enriched for three out of four xylogenesis-related datasets, while only a slight non-significant enrichment was observed for the EgrMYB-gene network that excluded binding events in accessible chromatin (Fig. 7). No such enrichment was observed for the HaloTag-DHS-gene negative control or the non-SCW-related dataset, showing that the link between TF binding, accessible chromatin and target gene functional enrichment is specific to the EgrMYB network.

Discussion

The roles of noncoding DNA elements in regulating wood formation are poorly understood, partially due to the difficulty in identifying them. Accessible chromatin representing nucleosome-depleted regions have a strong association with preferential binding of regulatory proteins to CREs at promoter and enhancer regions (Felsenfeld *et al.*, 1996; Zhang, W. *et al.*, 2012; Zhang *et al.*, 2015; Zhu *et al.*, 2015). In plants, whole-genome identification of such DHSs has been mostly limited to seedlings, roots, leaves and flowers in *Arabidopsis thaliana*, seedling, ear and husk in *Zea mays,* developing fruit of tomato, vascular cambium in *Populus*, and seedling and callus in *Oryza sativa* (Zhang, W. *et al.*, 2012; Zhang, Wenli *et al.*, 2012; Gent *et al.*, 2014; Pajoro *et al.*, 2014; Sullivan *et al.*, 2014; Vera *et al.*, 2014; Cumbie *et al.*, 2015; Qiu *et al.*, 2016; Rodgers-Melnick *et al.*, 2016; Liu *et al.*, 2017; Oka *et al.*, 2017; Zinkgraf *et al.*, 2017; Alexandre *et al.*, 2018) , in addition to a comparative study of accessible chromatin across four herbaceous plants (Maher *et al.*, 2018). Cell type-specific chromatin accessibility assays have also recently been implemented in plants (Sijacic *et al.*, 2018; Tannenbaum *et al.*, 2018). Our study intended to provide the first comprehensive overview of accessible chromatin in developing xylem. It contributes high-confidence DHSs that are

conserved between clonal ramets as well as evidence of their importance to *Eucalyptus* gene regulation and EgrMYB-family transcriptional networks linked to systems biology studies of wood biology.

Vierstra *et al.* (2014) found that DNase-seq fragment sizes reflect the chromatin structure in which resulting DHSs are identified, specifically those that are depleted of nucleosomes (small DNase-seq fragments) and those that span nucleosomes in a loose chromatin configuration (large DNase-seq fragments). The library size selection used in this study clearly enriched the small-fragment library for TSS-proximal DHSs associated with TF-binding, given the higher enrichment for gene regions, TSSs and CNSs compared to the large-fragment DHSs (Fig. 2b, Table S5). Our two-pronged approach hence distinguishes these overlapping but differing chromatin contexts in immature xylem.

We demonstrated that, in a broad sense, IX-DHS represent functional regions of the (mostly) noncoding genome due to their significant association with, or proximity to, genes and TSSs, gene expression, CNSs, depleted SNV density, and histone modifications. The tight relationship between IX-DHSs, SNV depletion and CNS enrichment observed in this study is likely due to purifying selection at protein binding sites (Blanchette & Tompa, 2002; Haudry *et al.*, 2013; Korkuć *et al.*, 2014). In human populations, for example, DHSs have reduced SNP diversity overall (The ENCODE Project Consortium, 2012). In one study, CNSs identified in *A. thaliana* overlapped flower and leaf DHSs by a fold-enrichment of 4.0 (Van de Velde *et al.*, 2014), a degree very similar to what we observed with IX-DHSs (Table

S5). Enrichment of IX-DHSs with modified histones is also significant because, at an epigenetic level, modified histones act as "signatures" of enhancer or promoter CREs regulating gene expression (reviewed by Weber *et al.*, 2016). H3K4me3, for example, can distinguish CREs near TSSs, while H3K4me1 can identify activated (H3K4me1-enriched) or inactive/poised (H3K27me3-enriched) distal enhancers occurring in accessible chromatin in mammals (Heintzman *et al.*, 2007; Rada-Iglesias *et al.*, 2011). The enrichment of IX-DHSs for both H3K4me3 and H3K27me3 (Fig. 6), suggests that the DHSs identified are relevant to both proximal and distal regulatory elements. In maize seedlings and callus, an enrichment for H3K27me3 in close proximity to DHSs was observed only for intergenic regions (Zhang, W. *et al.*, 2012), which comprise the vast majority of IX-DHSs in our data (Fig. 3a). It may seem counterintuitive that H3K27me3 could flank certain DHSs: this association seems to be the case when a mixture of cell types containing cell type-specific DHSs are analyzed, since cell type-specific DHSs are enriched for H3K27me3 whereas ubiquitous DHSs are not (Shu *et al.*, 2011) and single-cell DHSs negatively correlate with H3K27me3 (Jin *et al.*, 2015). In some cases this could also be explained by bivalent nucleosomes containing both H3K4me3 and H3K27me3 (Chen *et al.*, 2018). While no H3K4me1 or H3K27ac data has yet been generated in *Eucalyptus*, histone modification co-localisation may be used as a second line of evidence to identify intergenic enhancer-exposing DHSs.

We also provided evidence that open chromatin regions in IX-DHS are of relevance to the regulation of genes underlying certain processes underlying wood

formation through EgrMYB TFs which are homologous to those regulating SCW deposition in *Arabidopsis* (Hussey *et al.*, 2019). The significant enrichment for accessible chromatin among *in vitro* binding sites for all 6 EgrMYB TFs suggests that many of them are functional and can be prioritised within the context of the immature xylem chromatin structure. In fact, DNase I footprints at well-characterized motifs have been used to infer transcriptional regulatory networks (Van de Velde *et al.*, 2014; Sullivan *et al.*, 2015). Based on this rationale, restricting EgrMYB binding events to those occurring in accessible chromatin clearly improved the enrichment of their potential target genes for several systems biology datasets linking co-expression modules and candidate genes to wood formation processes and/or lignocellulosic biomass traits (Fig. 7). A similar approach of linking MYB binding sites to putative targets through accessible chromatin was recently adopted by Maher *et al.* (2018), allowing them to identify a MYB regulatory module associated with cell fate and abiotic stress regulation. As a caveat, we acknowledge that using proximity of TF binding events to genes alone is a complex topic and does not necessarily imply a regulatory relationship between TFs and potential target genes. Furthermore, it cannot differentiate between positive and negative transcriptional regulatory effects. Thus, the protein-DNA networks reconstructed here will need to be functionally validated.

The number of DHSs identified in this study is higher than those of AP1-induced seedlings (up to 8,789; Pajoro *et al.*, 2014), somewhat less than the 38,290 leaf and 41,193 flower tissue DHSs identified in *A. thaliana* (Zhang, Wenli *et al.*, 2012), and much lower than 97,975 to 371,692 DHSs reported in *O. sativa* or *P. trichocarpa*

(Zhang, W. *et al.*, 2012; Zinkgraf *et al.*, 2017). A likely reason for higher numbers in some studies is the use of F-seq (Boyle, Alan P. *et al.*, 2008), which in our study yielded large numbers of DHSs using default settings (Fig. S3a), coupled to statistical approaches that control for technical rather than biological false discovery rates. The critical assessment of DNase-seq analysis algorithms and biological reproducibility adopted in this study allowed us to identify high-confidence, albeit it stringent, DHSs in immature xylem. Another strength of our study was the use of a naked-DNA control library to account for intrinsic DNase-seq bias evident in its significant association of NDC DHSs with expression (Fig. 4a) and CNSs (Table S3). A further advantage of our experimental design is that the modified histone and RNA pol II ChIP-seq, RNA-seq and DNase-seq data all originated from the same samples, allowing for a direct comparison of overlap metrics and enrichment in these tissues.

DNase-seq profiles from other *E. grandis* tissues will help to identify tissue-specific DHSs that serve critical regulatory roles in wood formation specifically. In our experience, each tissue poses unique optimization challenges for DNase-seq, with some failing to produce high-quality DNase-seq data using the methods followed here possibly due to a high phenolics content (not shown). Linking distal DHSs (housing potential enhancers) to target genes will greatly benefit from Hi-C (Lieberman-Aiden *et al.*, 2009) or other chromatin interaction data for developing wood, a feat that has not yet been achieved. Nonetheless, DNase-seq data from this study provides critical information on the chromatin level of gene expression, serving as a useful addition to integrative multi-omics approaches aimed at

unravelling noncoding genome function and gene regulatory network architecture.

Acknowledgements

We thank Karen van der Merwe (University of Pretoria) and the Centre for High Performance Computing (www.chpc.ac.za) for bioinformatics assistance and infrastructure. We acknowledge funding from the National Research Foundation (NRF) of South Africa - Bioinformatics and Functional Genomics Programme (BFG Grant UID 86936 and 97911), the Department of Science and Technology (Strategic Grant for the *Eucalyptus* Genomics Platform), the Technology and Human Resources for Industry Programme (Grant UID 80118 and 96413), Mondi Ltd and Sappi Ltd. LT, KB and MO acknowledge postgraduate scholarship from the NRF. We thank Jonathan Featherston and Dirk Swanevelder (Agricultural Research Council, Pretoria) for assistance with Covaris sonication.

Author contributions

KB performed the experimental work and co-wrote the manuscript with SH. SH, EM and AM supervised the study and revised the manuscript. LT performed the DAP-seq experiments and reviewed the manuscript. MO conducted the whole-genome re-sequencing and SNV imputation analysis and reviewed the manuscript.

Additional information

Accession codes

The sequence data generated in this study was submitted to NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/), BioProject accessions PRJNA504747, PRJNA505134 and PRJNA505767. DHS regions can be visualized and downloaded from the JBrowse instance (for genome assembly v.1.1 and v.2.0) at http://eucgenie.org.

References

Alexandre CM, Urton JR, Jean-Baptiste K, Huddleston J, Dorrity MW, Cuperus JT, Sullivan AM, Bemm F, Jolic D, Arsovski AA, et al. 2018. Complex relationships between chromatin accessibility, sequence divergence, and gene expression in *Arabidopsis thaliana*. *Mol Biol Evol* **35**(4): 837-854.

Andrews S 2010. FastQC: a quality control tool for high throughput sequence data.

Banf M, Rhee SY. 2017. Computational inference of gene regulatory networks: Approaches, limitations and opportunities. *Biochim Biophys Acta* **1860**(1): 41-52.

Bartlett A, O'Malley RC, Huang SC, Galli M, Nery JR, Gallavotti A, Ecker JR. 2017. Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat Protoc* **12**(8): 1659-1672.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**: 289.

Bhargava A, Ahad A, Wang S, Mansfield SD, Haughn GW, Douglas CJ, Ellis BE. 2013. The interacting MYB75 and KNAT7 transcription factors modulate

secondary cell wall deposition both in stems and seed coat in *Arabidopsis*. *Planta* **237**: 1199-1211.

Blanchette M, Tompa M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res* **12**(5): 739-748.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2): 311-322.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**(21): 2537-2538.

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* **21**(3): 456-464.

Brondani RP, Brondani C, Grattapaglia D. 2002. Towards a genus-wide reference linkage map for *Eucalyptus* based exclusively on highly informative microsatellite markers. *Mol Genet Genomics* **267**(3): 338-347.

Brondani RPV, Brondani C, Tarchini R, Grattapaglia D. 1998. Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. *TAG Theoretical and Applied Genetics* **97**(5-6): 816-827.

Byrne M, Marquezgarcia MI, Uren T, Smith DS, Moran GF. 1996. Conservation and Genetic Diversity of Microsatellite loci in the Genus *Eucalyptus*. *Aust. J. Bot.* **44**(3).

Calo E, Wysocka J. 2013. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**(5): 825-837.

Chen A, Chen D, Chen Y. 2018. Advances of DNase-seq for mapping active gene regulatory elements across the genome in animals. *Gene* **667**: 83-94.

Chen H, Wang JP, Liu H, Li H, Lin YJ, Shi R, Yang C, Gao J, Zhou C, Li Q, et al. 2019. Hierarchical transcription-factor and chromatin binding network for wood formation in *Populus trichocarpa*. *Plant Cell*.

Cumbie JS, Filichkin SA, Megraw M. 2015. Improved DNase-seq protocol facilitates high resolution mapping of DNase I hypersensitive sites in roots in *Arabidopsis thaliana*. *Plant Methods* **11**: 42.

Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. 2017. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Res* **27**(12): 2096-2107.

Dharanishanthi V, Dasgupta MG. 2016. Construction of co-expression network based on natural expression variation of xylogenesis-related transcripts in *Eucalyptus tereticornis*. *Mol Biol Rep* **43**(10): 1129-1146.

Du Z, Li H, Wei Q, Zhao X, Wang C, Zhu Q, Yi X, Xu W, Liu XS, Jin W, et al. 2013. Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in *Oryza sativa* L. Japonica. *Mol Plant* **6**(5): 1463-1472.

Engelhorn J, Blanvillain R, Kröner C, Parrinello H, Rohmer M, Posé D, Ott F, Schmid M, Carles C. 2017. Dynamics of H3K4me3 chromatin marks prevails over H3K27me3 for gene regulation during flower morphogenesis in *Arabidopsis thaliana*. *Epigenomes* **1**(2).

Felsenfeld G, Boyes J, Chung J, Clark D, Studitsky V. 1996. Chromatin structure and gene expression. *Proceedings of the National Academy of Sciences* **93**(18): 9384-9388.

Galas DJ, Schmitz A. 1978. DNase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research* **5**(9): 3157-3170.

Gent JI, Madzima TF, Bader R, Kent MR, Zhang X, Stam M, McGinnis KM, Dawe RK. 2014. Accessible DNA and relative depletion of H3K9me2 at maize loci undergoing RNA-directed DNA methylation. *Plant Cell* **26**(12): 4903-4917.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**(6012): 1775-1787.

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**(8): R86.

Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57**: 159-197.

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**(8): e1002638.

Hardison R. 2000. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends in Genetics* **16**(9): 369-372.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* **45**(8): 891-898.

He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H, et al. 2014. Refined DNase-seq protocol and data analysis reveals intrinsic

bias in transcription factor footprint identification. *Nat Methods* **11**(1): 73-78.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**(3): 311-318.

Hussey S, Grima-Pettenati J, Myburg AA, Mizrachi E, Brady SM, Yoshikuni Y, Deutsch S. 2019. A standardized synthetic *Eucalyptus* transcription factor and promoter panel for re-engineering secondary cell wall regulation in biomass and bioenergy crops. *ACS Synth Biol* **8**(2): 463-465.

Hussey SG, Loots MT, van der Merwe K, Mizrachi E, Myburg AA. 2017. Integrated analysis and transcript abundance modelling of H3K4me3 and H3K27me3 in developing secondary xylem. *Sci Rep* **7**(1): 3370.

Hussey SG, Mizrachi E, Creux NM, Myburg AA. 2013. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Front Plant Sci* **4**: 325.

Hussey SG, Mizrachi E, Groover A, Berger DK, Myburg AA. 2015. Genome-wide mapping of histone H3 lysine 4 trimethylation in *Eucalyptus grandis* developing xylem. *BMC Plant Biology* **15**: 117.

Jiang J. 2015. The 'dark matter' in the plant genomes: non-coding and unannotated DNA sequences associated with open chromatin. *Current Opinion in Plant Biology* **24**: 17-23.

Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, Ni B, Sklar J, Przytycka TM, Childs R, et al. 2015. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**(7580): 142-146.

Kalluri UC, Yin H, Yang X, Davison BH. 2014. Systems and synthetic biology approaches to alter plant cell walls and reduce biomass recalcitrance. *Plant Biotechnol J* **12**(9): 1207-1216.

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**(6): 1099-1107.

Koohy H, Down TA, Spivakov M, Hubbard T. 2014. A comparison of peak callers used for DNase-Seq data. *PLoS ONE* **9**(5): e96303.

Korkuć P, Schippers JH, Walther D. 2014. Characterization and identification of cis-regulatory elements in *Arabidopsis* based on single-nucleotide polymorphism information. *Plant Physiol* **164**(1): 181-200.

Lafon-Placette C, Faivre-Rampant P, Delaunay A, Street N, Brignolas F, Maury S. 2013. Methylome of DNase I sensitive chromatin in *Populus trichocarpa* shoot apical meristematic cells: a simplified approach revealing characteristics of gene-body DNA methylation in open chromatin state. *New Phytol* **197**(2): 416-430.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Lazarovici A, Zhou T, Shafer A, Dantas Machado AC, Riley TR, Sandstrom R, Sabo PJ, Lu Y, Rohs R, Stamatoyannopoulos JA, et al. 2013. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc Natl Acad Sci U S A* **110**(16): 6376-6381.

Li E, Wang S, Liu Y, Chen JG, Douglas CJ. 2011. OVATE FAMILY PROTEIN4 (OFP4) interaction with KNAT7 regulates secondary cell wall formation in *Arabidopsis thaliana*. *Plant J* **67**(2): 328-341.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5): 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li X, Wang X, He K, Ma Y, Su N, He H, Stolc V, Tongprasit W, Jin W, Jiang J, et al. 2008. High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *The Plant Cell* **20**: 259-276.

Li Z, Fernie AR, Persson S. 2016. Transition of primary to secondary cell wall synthesis. *Science Bulletin* **61**(11): 838-846.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950): 289-293.

Liu Y, Zhang W, Zhang K, You Q, Yan H, Jiao Y, Jiang J, Xu W, Su Z. 2017. Genome-wide mapping of DNase I hypersensitive sites reveals chromatin accessibility changes in *Arabidopsis* euchromatin and heterochromatin regions under extended darkness. *Sci Rep* **7**(1): 4093.

Lockton S, Gaut B. 2005. Plant conserved non-coding sequences and paralogue evolution. *Trends in Genetics* **21**: 60-65.

Lu Z, Ricci WA, Schmitz RJ, Zhang X. 2018. Identification of cis-regulatory elements by chromatin structure. *Curr Opin Plant Biol* **42**: 90-94.

Maher KA, Bajic M, Kajala K, Reynoso M, Pauluzzi G, West DA, Zumstein K, Woodhouse M, Bubb K, Dorrity MW, et al. 2018. Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell* **30**(1): 15-36.

McCarthy RL, Zhong R, Ye Z-H. 2009. MYB83 is a direct target of SND1 and acts redundantly with MYB46 in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. *Plant and Cell Physiology* **50**(11): 1950-1964.

Meyer CA, Liu XS. 2014. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* **15**(11): 709-721.

Mitsuda N, Iwase A, Yamamoto H, Yoshida M, Seki M, Shinozaki K, Ohme-Takagi M. 2007. NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of *Arabidopsis*. *Plant Cell* **19**(1): 270-280.

Mizrachi E, Verbeke L, Christie N, Fierro AC, Mansfield SD, Davis MF, Gjersing E, Tuskan GA, Van Montagu M, Van de Peer Y, et al. 2017. Network-based integration of systems genetics data reveals pathways associated with lignocellulosic biomass accumulation and processing. *Proc Natl Acad Sci U S A* **114**(5): 1195-1200.

Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, Jenkins J, Lindquist E, Tice H, Bauer D, et al. 2014. The genome of *Eucalyptus grandis*. *Nature* **510**(7505): 356-362.

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**: 288.

Narlikar L, Ovcharenko I. 2009. Identifying regulatory elements in eukaryotic genomes. *Brief Funct Genomic Proteomic* **8**(4): 215-230.

Oka R, Zicola J, Weber B, Anderson SN, Hodgman C, Gent JI, Wesselink JJ, Springer NM, Hoefsloot HCJ, Turck F, et al. 2017. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol* **18**(1): 137.

Pajoro A, Madrigal P, Muino JM, Matus JT, Jin J, Mecchia MA, Debernardi JM, Palatnik JF, Balazadeh S, Arif M, et al. 2014. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biol* **15**(3): R41.

Pinard D, Fierro AC, Marchal K, Myburg AA, Mizrachi E. 2019. Organellar carbon metabolism is co-ordinated with distinct developmental phases of secondary xylem. *New Phytol*.

Ployet R, Veneziano Labate MT, Regiani Cataldi T, Christina M, Morel M, San Clemente H, Denis M, Favreau B, Tomazello Filho M, Laclau JP, et al. 2019. A systems biology view of wood formation in *Eucalyptus grandis* trees submitted to different potassium and water regimes. *New Phytol*.

Qiu Z, Li R, Zhang S, Wang K, Xu M, Li J, Du Y, Yu H, Cui X. 2016. Identification of regulatory DNA elements using genome-wide mapping of DNase I hypersensitive sites during tomato fruit development. *Mol Plant* **9**(8): 1168-1182.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**(7333): 279-283.

Ritchie GR, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. *Nat Methods* **11**(3): 294-296.

Rodgers-Melnick E, Vera DL, Bass HW, Buckler ES. 2016. Open chromatin reveals the functional maize genome. *Proc Natl Acad Sci U S A* **113**(22): E3177-3184.

Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, Shafer A, Kawamoto J, Hall R, Mack J, Dorschner MO, et al. 2004. Discovery of functional

noncoding elements by digital analysis of chromatin structure. *Proc Natl Acad Sci U S A* **101**(48): 16837-16842.

Shu W, Chen H, Bo X, Wang S. 2011. Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic Acids Res* **39**(17): 7428-7443.

Sijacic P, Bajic M, McKinney EC, Meagher RB, Deal RB. 2018. Changes in chromatin accessibility between *Arabidopsis* stem cells and mesophyll cells illuminate cell type-specific transcription factor networks. *Plant J* **94**(2): 215-231.

Soler M, Camargo ELO, Carocha V, Cassan-Wang H, Clemente HS, 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.

Soler M, Plasencia A, Lepikson-Neto J, Camargo EL, 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**: 1422.

Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res* **21**(10): 1757-1767.

Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, Sandstrom R, Thurman RE, Neph S, Reynolds AP, et al. 2014. Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. *Cell Rep* **8**(6): 2015-2030.

Sullivan AM, Bubb KL, Sandstrom R, Stamatoyannopoulos JA, Queitsch C. 2015. DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. *Current Plant Biology* **3-4**: 40-47.

Tannenbaum M, Sarusi-Portuguez A, Krispil R, Schwartz M, Loza O, Benichou JIC, Mosquna A, Hakim O. 2018. Regulatory chromatin landscape in *Arabidopsis thaliana* roots uncovered by coupling INTACT and ATAC-seq. *Plant Methods* **14**: 113.

The Arabidopsis Interactome Mapping Consortium. 2011. Evidence for network evolution in an *Arabidopsis* interactome map. *Science* **333**: 601-607.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.

The modENCODE Consortium. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**(6012): 1787-1797.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75-82.

Van de Velde J, Heyndrickx KS, Vandepoele K. 2014. Inference of transcriptional networks in *Arabidopsis* through conserved noncoding sequence analysis. *The Plant Cell* **26**: 2729-2745.

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**(4): 2586-2598.

Vera DL, Madzima TF, Labonne JD, Alam MP, Hoffman GG, Girimurugan SB, Zhang J, McGinnis KM, Dennis JH, Bass HW. 2014. Differential nuclease sensitivity profiling of chromatin reveals biochemical footprints coupled to

gene expression and functional DNA elements in maize. *Plant Cell* **26**(10): 3883-3893.

Vierstra J, Wang H, John S, Sandstrom R, Stamatoyannopoulos JA. 2014. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nat Methods* **11**(1): 66-72.

Vining KJ, Romanel E, Jones RC, Klocko A, Alves-Ferreira M, Hefer CA, Amarasinghe V, Dharmawardhana P, Naithani S, Ranik M, et al. 2014. The floral transcriptome of *Eucalyptus grandis*. *New Phytologist* **206**(4): 1406-1422.

Wang Q, Ci D, Li T, Li P, Song Y, Chen J, Quan M, Zhou D, Zhang D. 2016. The role of DNA methylation in xylogenesis in different tissues of poplar. *Frontiers in Plant Science* **7**: 1003.

Weber B, Zicola J, Oka R, Stam M. 2016. Plant Enhancers: A Call for Discovery. *Trends Plant Sci* **21**(11): 974-987.

Weinhofer I, Hehenberger E, Roszak P, Hennig L, Köhler C. 2010. H3K27me3 profiling of the endosperm implies exclusion of polycomb group protein targeting by DNA methylation. *PLoS Genetics* **6**(10): e1001152.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*: Springer-Verlag New York.

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* **9**: 1535.

Zhang T, Zhang W, Jiang J. 2015. Genome-wide nucleosome occupancy and positioning and their impact on gene expression and evolution in plants. *Plant Physiol* **168**(4): 1406-1416.

Zhang W, Wu Y, Schnable JC, Zeng Z, Freeling M, Crawford GE, Jiang J. 2012. High-resolution mapping of open chromatin in the rice genome. *Genome Res* **22**(1): 151-162.

Zhang W, Zhang T, Wu Y, Jiang J. 2012. Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. *The Plant Cell* **24**: 2719-2731.

Zhang X, Bernatavichute YV, Cokus S, Pellegrini M, Jacobsen SE. 2009. Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biol* **10**: R62.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, et al. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126**(6): 1189-1201.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**: R137.

Zhong R, Lee C, Ye Z-H. 2010. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in Plant Science* **15**(11): 625-632.

Zhong R, Lee C, Zhou J, McCarthy RL, Ye Z-H. 2008. A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. *The Plant Cell* **20**: 2763-2782.

Zhong R, McCarthy RL, Lee C, Ye Z-H. 2011. Dissection of the transcriptional program regulating secondary wall biosynthesis during wood formation in poplar. *Plant Physiology* **157**: 1452–1468.

Zhou J, Lee C, Zhong R, Ye Z-H. 2009. MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in *Arabidopsis*. *The Plant Cell* **21**(1): 248-266.

Zhu B, Zhang W, Zhang T, Liu B, Jiang J. 2015. Genome-wide prediction and validation of intergenic enhancers in Arabidopsis using open chromatin signatures. *Plant Cell* **27**(9): 2415-2426.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics* **39**(1): 61-69.

Zinkgraf M, Liu L, Groover A, Filkov V. 2017. Identifying gene coexpression networks underlying the dynamic regulation of wood-forming tissues in *Populus* under diverse environmental conditions. *New Phytol* **214**(4): 1464-1478.