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

Authors: Katrien Brown, Lazarus T. Takawira, Marja M. O'Neill, Eshchar Mizrachi, Alexander A. Myburg and Steven G. Hussey

Article acceptance date: 29 April 2019

The following **Supporting Information** is available for this article:

Fig. S1 Agarose gel electrophoresis images of immature xylem DNase-seq library quality-control and preparation.

Fig. S2 Performance of various peak-calling algorithms in identifying DNase I hypersensitive sites (DHSs).

Fig. S3 Irreproducible discovery rate plots for immature xylem DNase-seq data.

Fig. S4 Absolute expression levels of genes overlapping immature xylem DNase I hypersensitive sites in seven *Eucalyptus* tissues and organs.

Fig. S5 Proximal enrichment of small-fragment, immature xylem (pooled-fragment) and large-fragment DNase I hypersensitive sites to H3K4me3, H3K27me3 and transcriptional start sites.

Fig. S6 Degree distributions of nodes in transcription factor-target gene networks involving EgrMYB transcription factors.

Table S1 Parameters for sequence read mapping and variant detection for *E. grandis* TAG0014 reference genome imputation.

Table S2 Immature xylem DNase-seq mapping rates.

Table S3 Summary of peak-calling algorithms tested.

Table S4 Biological reproducibility of immature xylem DNase I hypersensitive sites.

Methods S1 DNase I treatment, DNA isolation and sequencing

Note S1 Irreproducible discovery rate analysis.

Fig. S1. Agarose gel electrophoresis images of immature xylem DNase-seq library quality-control and preparation. (a) Gel separation of 50 bp marker (M), undigested chromatin (0 U) and chromatin treated with increasing units of DNase I (15 U, 25 U and 40 U) to assess undigested chromatin quality and optimize chromatin digestion. Small-fragment libraries were prepared from excised fragments ranging from 50 bp to <150 bp (1); large-fragment libraries were prepared by excision of fragments ranging 150 bp to 300 bp (2). **(b)** Separation of naked genomic DNA treated with 0 U, 7 mU, 10 U and 13 mU DNase I.

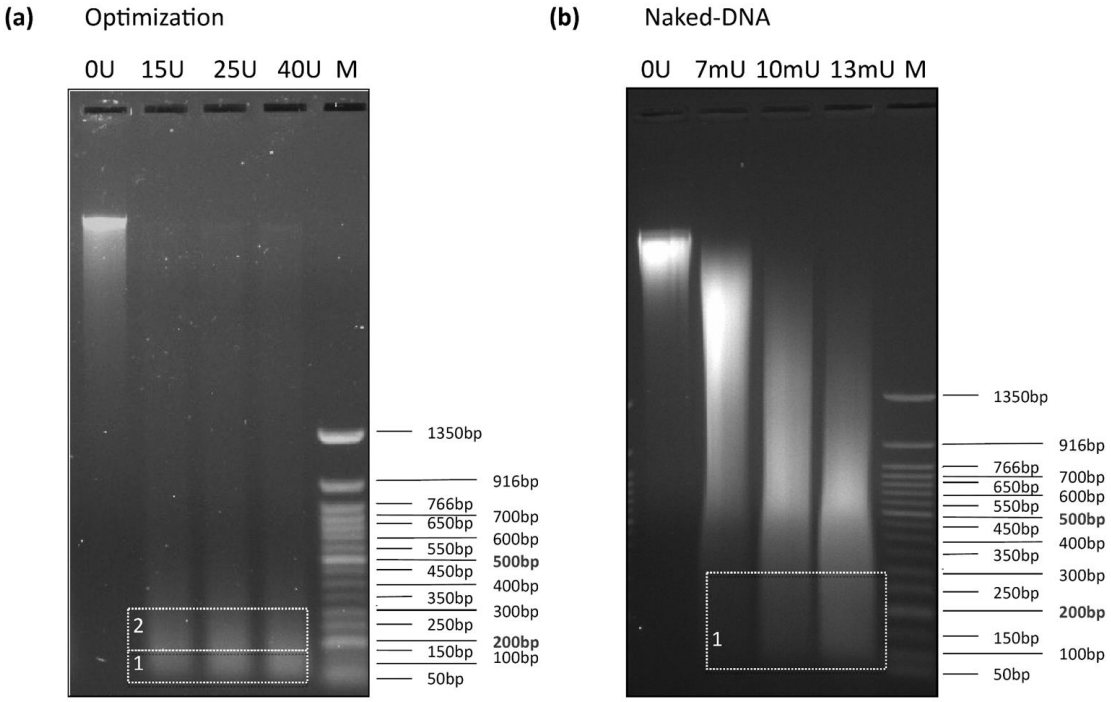

Fig. S2. Performance of various peak-calling algorithms in identifying DNase I hypersensitive sites (DHSs). (a) Number of DHS peaks identified by F-seq, Hotspot2 and MACS2 using a lenient threshold (FDR or *P*-value < 0.05; for Hotspot2 an FDR < 0.8 was used to obtain sufficient peaks for comparison) versus a stringent threshold (IDR < 0.05). Values are representative of the highest value from three replicates. **(b)** Biological reproducibility of DHSs identified by various peak-calling algorithms, expressed as Jaccard Index values of each pairwise biological replicate comparison. Only the small-fragment libraries were analyzed.

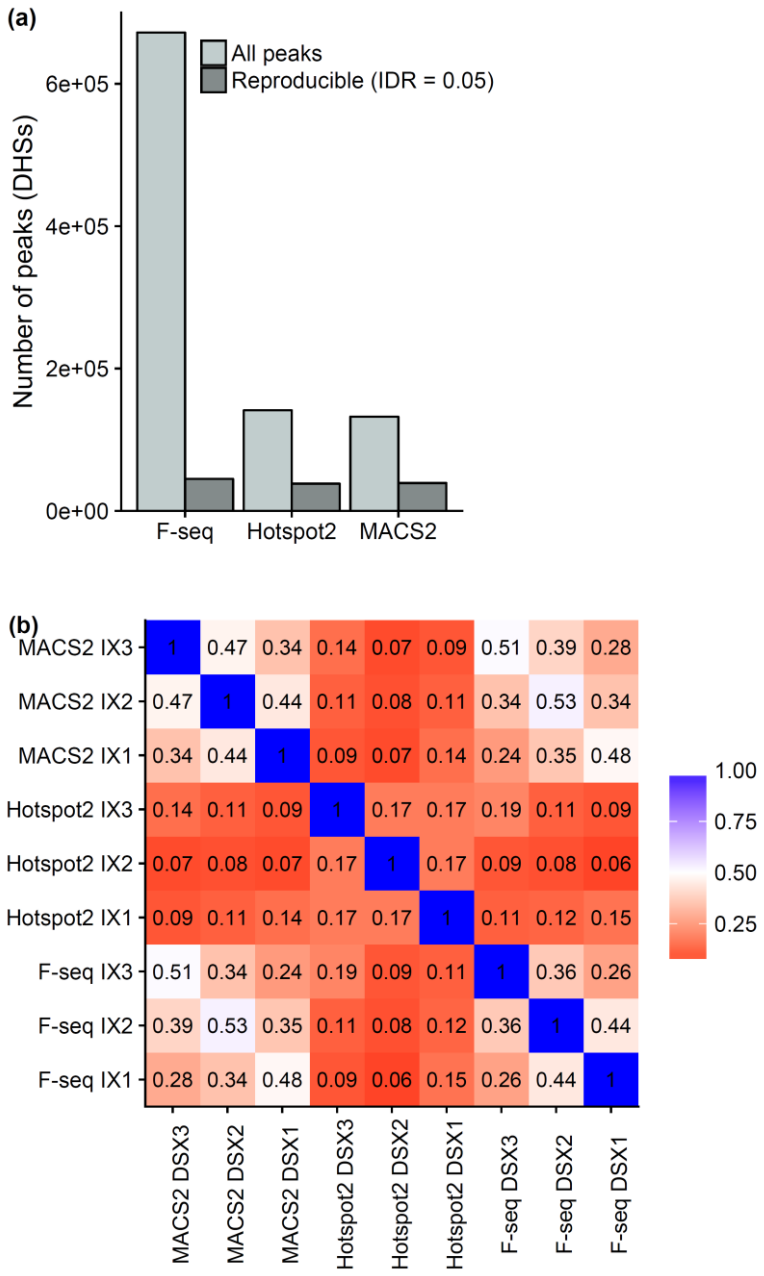

Fig. S. Irreproducible discovery rate plots for immature xylem DNase-seq data. The number of significant peaks that can be reproducibly called at a given IDR between **(a)** biological replicates samples from large-fragment libraries and **(b)** small-fragment libraries. Lines marked with an asterisk (*) resulted from a comparison with a discarded sample.

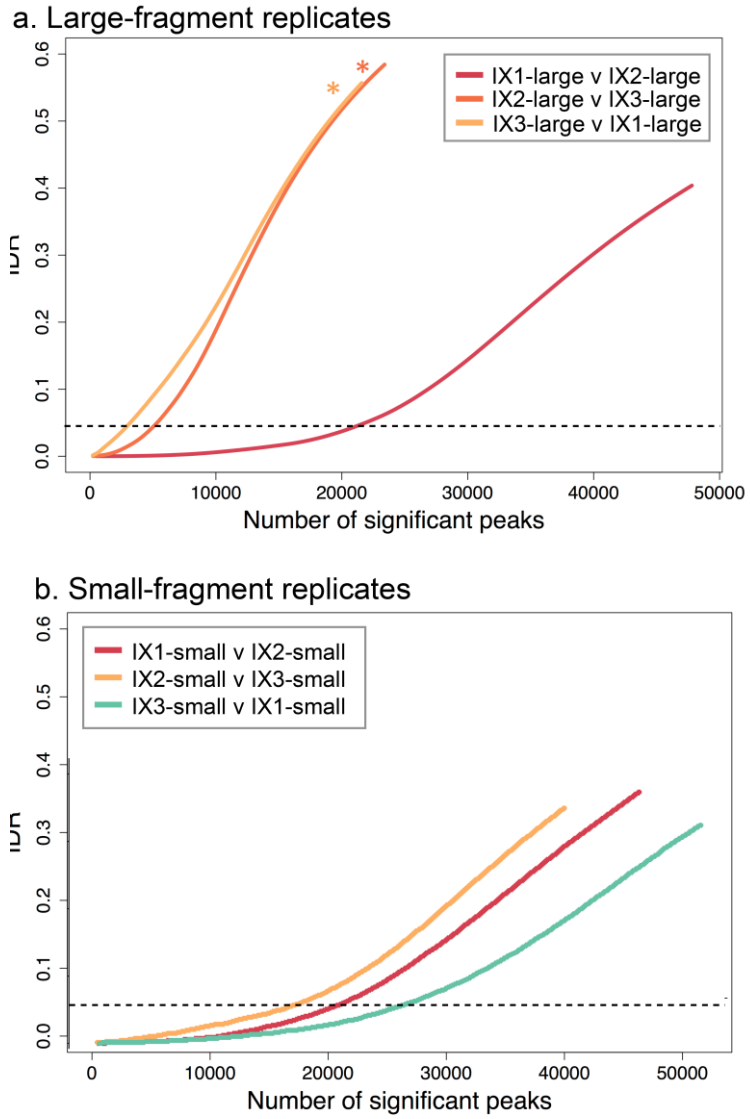

Fig. S4. Absolute expression levels of genes overlapping immature xylem DNase I hypersensitive sites in seven *Eucalyptus* tissues and organs. The kernel density estimation is indicated on the y-axis for each plot.

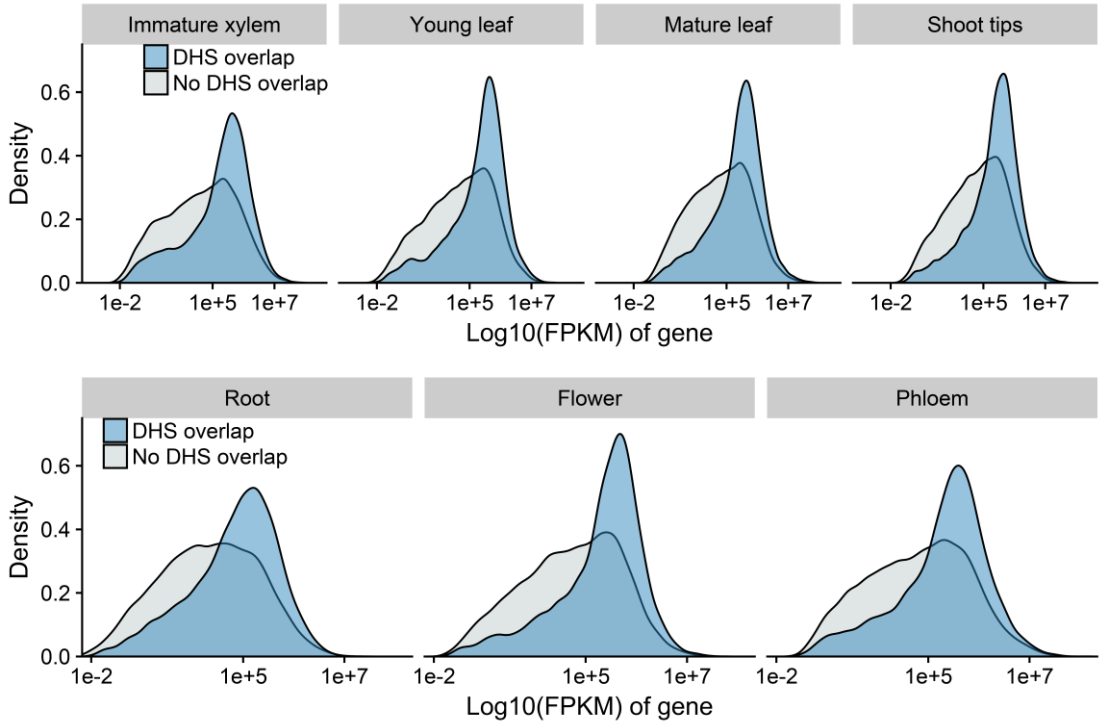

Fig. S5. Proximal enrichment of small-fragment, immature xylem (pooled-fragment) and large-fragment DNase I hypersensitive sites to H3K4me3, H3K27me3 and transcriptional start sites. Statistical significance assessed according to Fisher's Exact Test, using the median of 1000 random permutations as the null distribution.

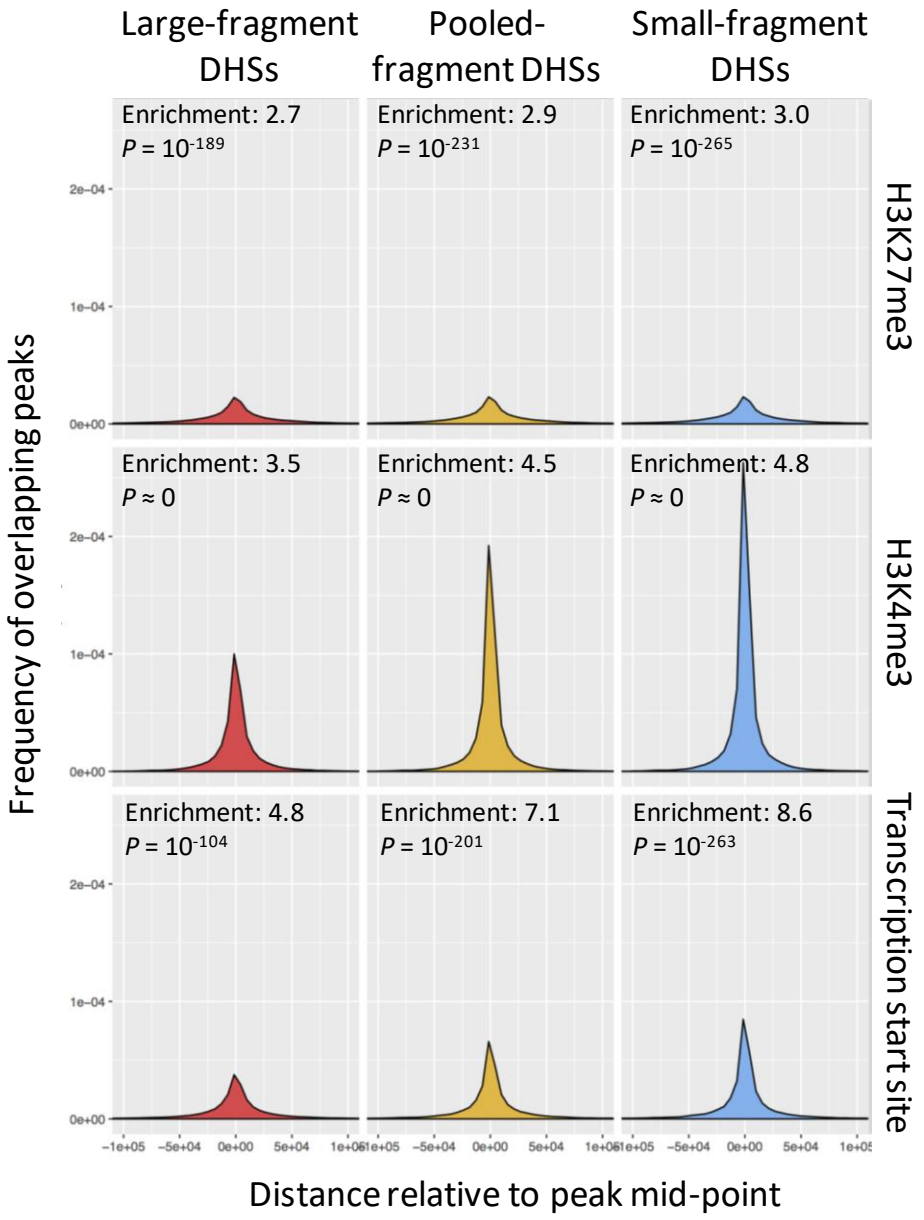

Fig. S6. Degree distributions of nodes in transcription factor-target gene networks involving EgrMYB transcription factors. (a) Degree distribution of the EgrMYB-DHS-gene network. **(b)** Degree distribution of the EgrMYB-non-DHS-gene network. n, number of nodes in the network; *k*, estimated exponent of the fitted power-law distribution.

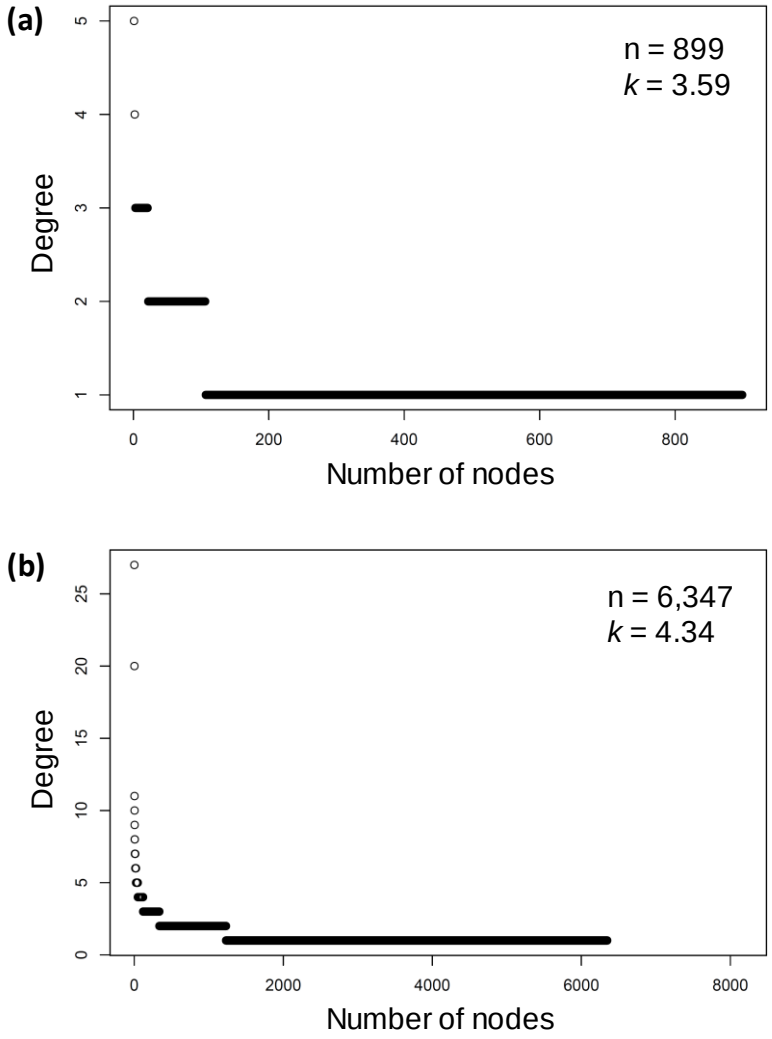

Table S1. Parameters for sequence read mapping and variant detection for *E. grandis* TAG0014 reference genome imputation.

Input parameter	Value
Map with BWA for Illumina (Li & Durbin, 2010)	
Reference genome	Egrandis_201.fa
Maximum edit distance	0
Fraction of missing alignments given 2% uniform base error rate	0.04
Maximum number of gap opens	1
Maximum number of gap extensions	-1
Disallow long deletion within 16 bp towards 3'-end	
Disallow InDel within 5 bp towards the end	
Number of first subsequences to take as seed	-1
Maximum edit distance in the seed	2
Mismatch penalty	3
Gap open penalty	11
Gap extension penalty	4
Iterative search enabled	true
Maximum number of alignments to output in the XA tag for reads paired properly	3
Maximum number of alignments to output in the XA tag for disconcordant read pairs (excluding singletons)	10
Maximum insert size for read pair to be considered as being mapped properly	350
Maximum occurrences for a read for pairing	100 000
MPileup (SAMtools)	
Reference genome	Egrandis_201.fa
Genotype Likelihood Computation performed	true
InDel calling performed	true
Phred-scaled gap extension sequencing error probability	20
Coefficient for modelling homopolymer errors	100
Skip InDel calling if the average per-sample depth is above	250
Phred-scaled gap open sequencing probability	40
Skip anomalous read pairs in variant calling	true
Enable probabilistic realignment for the computation of base alignment quality (BAQ)	true
Coefficient for downgrading mapping quality for reads containing excessive mismatches	0
Max reads per BAM	250
Extended BAQ computation	false
Minimum mapping quality for an alignment to be used	0
Minimum base quality for a base to be considered	13
Variant filtering using BCFtools (SAMtools)	
Minimum Root Mean Square (RMS) mapping quality for a SNP	10
Minimum read depth	5
Minimum number of alternate bases	2
SNPs within 3 bases around a gap were filtered	
Window size for filtering adjacent gaps	10
Minimum P-value for strand bias given PV4	0.0001
Minimum P-value for baseQ bias	1e-100
Minimum P-value for mapQ bias	0.0
Minimum P-value for end distance bias	0.0001
Minimum P-value for Hardy-Weinberg Equilibrium (plus F smaller than zero) bias	0.0001

Table S2. Immature xylem DNase-seq mapping rates.

Set ID	Sequences generated	Mapped reads	Duplicate reads	Mapping rate (%)	TagAlign reads*
IX1-small	43,382,931	35,906,338	2,012,641	82.8%	18,510,566
IX1-large	39,083,910	32,906,468	1,638,255	84.2%	15,775,510
IX2-small	36,195,973	29,840,076	1,595,568	82.4%	13,842,309
IX2-large	39,885,895	32,936,407	1,702,348	82.6%	14,172,772
IX3-small	37,482,967	30,709,603	1,920,475	81.9%	12,961,178
IX3-large	42,445,412	36,514,537	1,452,603	86.0%	19,398,925
NDC	49,239,963	42,725,028	3,517,759	86.8%	19,320,636
*WGR***	58,181,247	45,605,874	2,673,618	78.4%	23,711,132

NDC - Naked-DNA control; WGR - Whole-genome re-sequenced data

*The number of reads in the TagAlign format used for MACS2 and F-seq analyses. Mapping quality < 30 and all duplicates removed.

**Number of single-end reads selected from the paired-end immature xylem sample library.

Table S3. Summary of peak-calling algorithms tested.

Peak caller	Statistical method	Data metrics	Algorithm	Input control	Read shift
F-seq	FDR	Signal value	Kernel-based	No	No
Hotspot2	FDR	Signal value	Hotspot	No	No
MACS2	*P*-value or q-value	Signal value against background control	Dynamic Poisson distribution	Yes	Yes

Table S4. Biological reproducibility of immature xylem DNase I hypersensitive sites.

	Comparison	Number of peaks (IDR < 0.05)
Large-fragment replicates	IX1-large vs IX2-large	22,026
	IX2-large vs IX3-large	3,169*
	IX3-large vs IX1-large	5,139*
Small-fragment replicates	IX1-small vs IX2-small	19,941
	IX2-small vs IX3-small	25,319**
	IX3-small vs IX1-small	16,323

*Comparisons for which reproducibility was significantly lower (See methods) other comparisons and the DHSs sets were flagged as unreliable

** The highest number of DHSs that can be reproducibly called at IDR of 0.05

Table S5. Enrichment of conserved noncoding sequences (CNS) among various DNase I hypersensitive sites (DHS) datasets

	Overlapping CNSs	Overlapped DHSs	Enrichment*	*P*-value
Small-fragment DHS	14,420	5,046	4.80	≈ 0
Large-fragment DHS	8,823	3,487	3.64	≈ 0
Immature xylem DHSs	12,933	4,567	4.45	≈ 0
Naked-DNA DHSs	1,414	1,329	1.30	1.76×10^{-10}
Shuffled DHSs	1,026	1,026	1.00	NA

*****Enrichment is the ratio between DHSs overlapped by at least one CNS and the number of shuffled DHSs overlapped by at least one CNS (median value of 1000 permutations)
NA, Not Applicable

Methods S1

DNase I treatment, DNA isolation and sequencing

Through extensive optimisation, we found that immature xylem chromatin digested with 15U, 25U and 40U of RNase-free recombinant bovine pancrease DNase I (Roche, Basel, Switzerland) yielded a range of fragment sizes corresponding to the target DNase-seq library inserts (Fig. S1a). Extracted nuclei aliquots were kept cool on ice while the appropriate amounts of DNase I were added and mixed by tapping. Aliquots were then flash frozen in liquid nitrogen for 30 to 60 seconds before being placed into an incubator at 37 ºC with shaking at 700 rpm to thaw for 5 min in order to permeate the nuclear membrane. The reactions were then left to incubate for 10 minutes before the reaction was stopped using 500 ul of 50 mM EDTA buffer (pH 8). The solution was immediately cooled on ice before being centrifuged at 1100 x *g* for 10 min to pellet the DNase I treated nuclei.

DNA was extracted using commercial kits, with modifications. Briefly, 300 ul PL2 buffer from the Nucleospin Plant II DNA Extraction kit (Macherey-Nagel) was used to lyse the nuclei along with 10 ul RNase A. After mixing, the solution was incubated at 65 °C for 10 min followed by the addition of 75 ul PL3 and incubation on ice for 5 min. Debris was pelleted by centrifugation at 11,000 x *g* for 5 min and the supernatant passed through the filtration column by centrifugation at 11,000 x *g* for 2 min. Next, 200 ul NTI buffer (Nucleospin Gel and PCR Clean-up kit) per 100 ul lysate and an additional 400 ul ice-cold isopropanol (to aid with precipitation of small DNA fragments) was added and thoroughly mixed by gentle inversion of the tubes. The

mixture was incubated for 2 min on ice before being passed through the DNA binding column (Nucleospin Gel and PCR Clean-up kit) 600 ul at a time by centrifugation at 11,000 x *g* for 30 sec. The column was washed with 400 ul PW2 and 100 ul ice-cold 100% ethanol after 2 min incubation and centrifuged at maximum speed for 2 min to ensure the removal of residual ethanol before elution step. DNA was eluted using 35 ul of PE buffer.

Purified DNA from the digested samples was separated on a 1% agarose gel for library preparation. For the small-fragment library, DNA fragments ranging from 50 bp to <150 bp across all DNase concentrations were used and >150 bp to 300 bp for the large-fragment library (Fig. S1a). DNA was purified using the Nucleospin Gel and PCR Clean-up kit with the modifications mentioned above. Briefly, 200 ul NTI buffer (Gel and PCR Clean-up kit) per 100 mg gel was added and incubated at 50 °C until gel pieces were completely dissolved. Ice-cold isopropanol (200 ul) was added and the mixture was incubated for 2 min on ice before being passed through the DNA binding column 600 ul at a time by centrifugation at 11,000 x *g* for 30 sec. The column bound DNA was then washed with 400 ul PW2 and 100 ul ice-cold 100% ethanol after 2 min incubation. The column was centrifuged at maximum speed for 2 min to ensure the removal of residual ethanol before elution step. DNA was eluted using 35 ul of PE buffer in three parts, with 5 min incubation at room temperature between each elution (centrifuged at 11,000 x *g* for 1 min). Libraries were constructed by Novogene, Inc. (USA) and SE50 sequencing performed on the HiSeq2500 platform.

Note S1. **Irreproducible discovery rate analysis.**

It is a convention with the irreproducible discovery rate (IDR) method (Li *et al.*, 2011) implemented for epigenomic marks in the ENCODE project (Landt *et al.*, 2012) to identify an experiment-wide statistical threshold of biologically reproducible peaks to a peak set generated after bulking the individual biological replicates and ranking them by *P*-value. The IDR analysis begins with replicate-specific sets of peaks called under a lenient threshold, ranked by *P*-value. A bivariate rank distribution is then employed to quantify peak reproducibility between replicates based on physical peak overlap. The minimum number of biological replicates accepted by the ENCODE Consortium is two (Landt *et al.*, 2012). The authors additionally recommend that any additional biological replicates should have similar numbers of reproducible peaks (at least within a factor of 2 of each other). We implemented this approach for assessing biological reproducibility and identifying an experiment-wide statistical threshold for DNase I hypersensitivity sites. Comparing large-fragment libraries across biological replicates, peaks between the IX1-large and the IX2-large samples yielded 22,026 reproducible peaks (IDR < 0.05), while the reproducibility of IX3-large peaks with IX1-large and IX2-large was substantially lower by a factor of ~4 and 7, respectively, suggesting a poor quality dataset for sample IX3-large (Fig. S3a). We therefore discarded the IX3-large dataset since it did not meet the IDR criterion for additional biological replicates. For the small-fragment libraries, the maximum and minimum number of reproducible peaks (IDR < 0.05) was 25,319 (IX2-small vs IX3-small) and 16,323 (IX1-small vs IX3-small) respectively (Fig. S3, Table S4). This satisfied the biological reproducibility between all three replicates, and established the experiment-wide threshold of 25,319 peaks. Since this number is similar to the

number obtained for the IX-large data (22,026), we regarded this as an acceptable experiment-wide threshold for the three biological replicates.

References

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22**: 1813-1831.

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

Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat* **5**(3): 1752-1779.