Diversity and cis-element architecture of the promoter regions of cellulose synthase genes in *Eucalyptus*

Nicky M Creux*, Minique H De Castro*, Martin Ranik, Mathabatha F Maleka, Alexander A Myburg[§]

Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, Pretoria, 0028, South Africa

*These authors contributed equally to this work

[§]Corresponding author

Prof A.A. Myburg
Department of Genetics
University of Pretoria

Tel: +27 (0)12-4204945
Fax: +27 (0)12-3625327
Email: zander.myburg@fabi.up.ac.za

Email addresses:

NMC: nicky.creux@fabi.up.ac.za

MHDC: decastrom@arc.agric.za

MR: martin.ranik@gmail.com

MFM: malekamf@ufs.ac.za

AAM: zander.myburg@fabi.up.ac.za

Abstract

Lignocellulosic biomass from fast-growing plantation trees is composed of carbohydrate-rich materials deposited into plant cell walls in a coordinated manner during wood formation. The diversity and evolution of the transcriptional networks regulating this process have not been studied extensively. We investigated patterns of species-level nucleotide diversity in the

promoters of cellulose synthase (*CesA*) genes from different *Eucalyptus* tree species and assessed the possible roles of DNA sequence polymorphism in the gain or loss of cis-elements harboured within the promoters. Promoter regions of three primary and three secondary cell wall-associated *CesA* genes were isolated from 13 *Eucalyptus* species and were analysed for nucleotide and cis-element diversity. Species-level nucleotide diversity (π) ranged from 0.014 to 0.068 and different *CesA* promoters exhibited distinct patterns of sequence conservation. A set of 22 putative cis-elements were mapped to the *CesA* promoters using *in silico* methods. Forty two percent of the mapped cis-element occurrences contained singleton polymorphisms which resulted in either the gain or loss of a cis-element in a particular *Eucalyptus* species. The promoters of *Eucalyptus CesA* genes contained regions that are highly conserved at the species (*Eucalyptus*) and genus (with *Arabidopsis* and *Populus*) level, suggesting the presence of regulatory modules imposing functional constraint on such regions. Nucleotide polymorphisms in the *CesA* promoters more frequently created new cis-element occurrences than disrupted existing cis-element occurrences, a process which may be important for the maintenance and evolution of cellulose gene regulation in plants.

Keywords: Cis-element conservation, promoter evolution, secondary cell wall, wood formation, *CesA*, woody biomass

Introduction

With the current worldwide focus on renewable energy production and carbon sequestration (Ragauskas et al. 2006; Piao et al. 2009), lignocellulosic biomass from fast-growing plantation trees is being targeted as a renewable source of carbon for biofuels and biomaterials (Regalbuto 2009; Rathmann et al. 2010). The bulk of this biomass is comprised

of cellulose, hemi-cellulose and lignin contained in the secondary cell walls of wood fibre cells (Gorshkova et al. 2012). The biochemical and structural complexity of wood is determined largely by the coordinated expression of hundreds of regulatory, structural and biosynthetic genes (Aspeborg et al. 2005; Mellerowicz and Sundberg 2008). Several transcription factor genes have already been identified as key regulatory components of these pathways in the model plant *Arabidopsis* (Zhong et al. 2008; Zhong et al. 2010) and in woody plants such as *Populus and Eucalyptus* (Hu et al. 2010; Legay et al. 2010; McCarthy et al. 2010). Despite the emerging understanding of the molecular machinery underlying major biosynthetic pathways in wood-forming tissues, the nature and evolution of transcriptional networks regulating these pathways are not well described in woody plants such as *Eucalyptus*.

Eucalyptus is a richly diverse genus (Brooker 2000), containing over 700 species of woody plants, many of which are rapid and prolific producers of cellulose-rich biomass. Some of these species and their fast-growing hybrids form the basis of the most widely cultivated hardwood plantation crop in the world (Eldridge et al. 1994; Grattapaglia et al. 2009). The genus is divided into several subgenera (Pryor and Johnson 1971; Steane et al. 1999; Brooker 2000), of which the subgenus *Symphyomyrtus* is the largest and most diverse containing most of the commercially grown eucalypt species (Eldridge et al. 1994). The high phenotypic diversity in the subgenus is reflected at the DNA sequence level with nucleotide diversity values ranging from moderately diverse ($\pi = 0.0186$) for *E. grandis* (Novaes et al. 2008) to highly diverse ($\pi = 0.063$) for *E. loxophleba* (with single nucleotide polymorphisms every 16 to 33 positions, Kulheim et al. 2009). These values are similar to the levels of nucleotide diversity reported in other outbred forest tree genera such as *Pinus* ($\pi = 0.01$-0.02) and *Populus* ($\pi = 0.005$-0.01) (Reviewed in Neale and Ingvarsson 2008). The high nucleotide diversity in *Eucalyptus* and the recently completed *E. grandis* reference genome sequence

(DOE JGI, http://www.phytozome.net) provide opportunities to investigate the evolution and diversity of regulatory networks underlying wood development.

In eukaryotes, gene regulatory networks comprise cis-acting sequence elements in promoters, trans-acting elements or transcription factors which bind to promoter sequences and the genes regulated by both kinds of factors. Cis-elements are often found clustered together in promoter regions where transcription factors can bind as hetero- or homodimers to modulate the transcription of the gene (Reviewed in Farnham 2009). Cis-regulatory elements are often shared by the promoters of co-expressed genes due to common trans-regulation (Vandepoele et al. 2009). Another feature of cis-element sequences is their conservation in orthologous promoters from different species and genera, in comparison to flanking non-coding sequences (Freeling and Subramaniam 2009). These unique features of cis-regulatory elements are employed in a host of computational algorithms for the *in silico* detection of cis-elements (Tompa et al. 2005; Das and Dai 2007; Wijaya et al. 2008; Kim et al. 2009). Software programs, such as FootPrinter (Blanchette and Tompa 2003; Fang and Blanchette 2006), PHYLONET (Than et al. 2008) and PhyloScan (Carmack et al. 2007) are based on phylogenetic footprinting algorithms, which use the evolutionary relationships among promoter sequences to identify conserved cis-regulatory motifs (Blanchette et al. 2002; Blanchette and Tompa 2002). Phylogenetic footprinting approaches have been used to identify cis-elements in a number of plant species including *Populus* and *Eucalyptus* (Creux et al. 2008; Shi et al. 2010).

While cellulose deposition during secondary cell wall formation has been extensively studied in herbaceous and woody plants (Reviewed in Taylor 2008; Popper et al. 2011; Gorshkova et al. 2012), the promoters and cis-element composition of the cellulose synthase (*CesA*) genes have been the focus of only a small number of studies (Creux et al. 2008; Lu et al. 2008; Wu et al. 2009). CESA proteins, encoded by the *CesA* gene family, form large

membrane-embedded complexes depositing cellulose into plant cell walls (Mutwil et al. 2008). In embryophytes, this gene family forms a distinct clade consisting of 8-18 members per species (Hamann et al. 2004; Roberts and Bushoven 2007; Kumar et al. 2009; Yin et al. 2009). In plants there are two distinct *CesA* gene expression groups, one associated with primary cell wall formation and the other with secondary cell wall formation (Taylor et al. 2004; Desprez et al. 2007). Independent functional analyses of the *CesA* genes in woody and herbaceous species have revealed that orthologous *CesA* genes are functionally conserved in diverse plant species (Tanaka et al. 2003; Taylor et al. 2003; Samuga and Joshi 2004; Ranik and Myburg 2006; Kumar et al. 2009). This conservation likely includes a set of shared cis-regulatory sequences and transcription factors since *CesA* gene expression profiles are also highly conserved across different plant genera (Burton et al. 2004; Ranik and Myburg 2006; Creux et al. 2008).

In this study we hypothesize that cis-regulatory elements will coincide with regions of lower species-level nucleotide diversity in the promoters of evolutionary distinct *Eucalyptus* tree species. In addition, we hypothesize that different sets of cis-elements are conserved in the promoters of primary or secondary cell wall-related *CesA* genes. The objectives were: i) to quantify and assess patterns of species-level nucleotide diversity in the promoters of six *Eucalyptus* cellulose synthase genes, ii) to assess the effects of nucleotide polymorphism on putative cis-element occurrences and iii) to identify putative cis-elements that are differentially conserved in promoters of primary and secondary cell wall-related *CesA* genes. This study is the first to characterize the species-level nucleotide diversity and cis-element architecture of *CesA* gene promoters in a plant genus and adds to our understanding of the transcriptional regulation of this important plant gene family.

Materials and Methods

Plant material and DNA isolation

Leaf material was obtained from *Eucalyptus* species conservation hedges established with seed collected from natural stands in Australia (CSIRO, Online Resource 1). High quality genomic DNA was extracted from leaf material using the DNeasy® Plant Mini Kit (Qiagen, Valencia, CA). Genomic DNA was isolated from a single tree from each of 13 eucalypt species, including *E. fastigata* (subgenus *Eucalyptus*), and 12 species of the commercially important subgenus *Symphyomyrtus* representing three sections, *Latoangulatae*, *Maidenaria* and *Exsertaria* (Online Resource 1). In addition, genomic DNA was also isolated from nineteen *E. urophylla* individuals originating from seed collected on seven Indonesian islands (Timor, Flores, Alor, Pantar, Adonara, Lomblen and Wetar), broadly representing the geographical range of the species (Payn et al. 2008).

Promoter isolation and sequencing

Primer Designer software (version 5, Scientific and Educational Software, Durham, NC) was used to design primers (Online Resource 2) for the amplification of gene and promoter regions based on previously published *E. grandis CesA* gene and promoter sequences (Ranik and Myburg 2006; Creux et al. 2008). For *CesA1*, PCR amplification of a single gene fragment from each of the 13 *Eucalyptus* species, extending from the promoter to the end of intron 1, was performed (Online Resource 2). Only the promoter regions (approximately 1 kb upstream of the ATG) were amplified for the other five *CesA* genes. PCR was performed in 20 µl reaction volumes with 30 ng of genomic DNA, 0.4 µM of each primer, 0.20 mM of each dNTP and 0.15 U of ExSel DNA polymerase (Supertherm) with proofreading capability (4-fold lower error rate than standard Taq polymerase, according to manufacturers) using the following conditions: 30 cycles of denaturation at 94°C for 30 seconds, annealing at 56°C for 30 seconds and primer extension at 72°C for 2 minutes. The amplified fragments were cloned

(InsT/Aclone, MBI Fermentas, Hanover, MD) and a cloned copy of each promoter fragment was sequenced using overlapping Sanger reads (Macrogen Inc.), representing a single allele of each promoter from each species (Genbank accession numbers: JN573683 - JN573751).

The chromatograms were visually inspected to check the sequence quality of each base. When double, conflicting bases were identified at the same site the clone was re-sequenced and re-analyzed until a consensus was reached. At least 1000 bp of upstream sequence was analysed for *CesA1* to *5*, while approximately 800 nucleotides of the *CesA7* promoter were used. The 5' upstream regions isolated from each gene started at most 25 bp upstream of the start codon (ATG) and contained the 5' UTR and a minimum of 500 bp of each promoter. Sequences were analysed from the ATG because the start codon of each gene has previously been experimentally verified (Ranik and Myburg 2006).

Orthologous gene and promoter sequences of *Arabidopsis* and *Populus*

Kumar et al. (2009) proposed a new phylogeny-based nomenclature for the *CesA* genes in poplar. Their naming convention allows for direct comparison of the *Arabidopsis* and *Populus CesA* genes (Online Resource 3). This change in nomenclature has not yet been applied to the *Eucalyptus CesA* genes therefore in this study we kept the naming convention used in the first study that reported the *Eucalyptus CesA* genes (Ranik and Myburg 2006). Online Resource 3 lists the *Eucalyptus CesA* genes and their orthologs in *Populus* and *Arabidopsis*. Promoter sequences of the *Arabidopsis thaliana CesA* orthologs (*AtCesAs 1-10*) were obtained from The *Arabidopsis* Information Resource (TAIR9, www.arabidopsis.org), and those of the *Populus trichocarpa* orthologs (PtiCesAs, Kumar et al. 2009) were obtained from the *Populus* Genome Browser (DOE JGI http://www.phytozome.net, Online Resource 3). The same set of promoter sequences were used as in the study by Creux et al (2008). Additionally, the orthologous gene sequences of *Eucalyptus CesA1, AtCesA8* (At4g18780) and *PtiCesA8-A* (Pti235238), were obtained from the same databases. To ensure that the

regions of the *Eucalyptus CesA1* gene and its orthologs could be compared, the promoter regions were trimmed to equal length (297 bp) in all orthologs. The first intron region was also trimmed (to 68 bp), resulting in an analysis region that included part of the proximal promoter, 5' UTR (untranslated region), exon 1 and part of intron 1 (Accession numbers JN573752 - JN573783).

DNA sequence analysis

After removal of vector sequences, DNA sequences of the *Eucalyptus CesA* promoters were assembled using the Vector NTI software package (version 9.1.0, Invitrogen). The sequences were aligned using the Clustal W (Thompson et al. 1994) function of BioEdit (version 7.0.9, Hall 1999). DNA sequence analysis and nucleotide diversity (π; Nei and Li 1979) and θ_W; (Watterson 1975) calculations were performed using DnaSP (DNA Sequence Polymorphism, version 4.50.3, Rozas et al. 2003). The distribution of nucleotide diversity in the sequences was graphically represented using per site sliding windows of π. Nucleotide diversity was represented by the average of a moving window of 50 bp for Figure 1 and a sliding window of 100 bp (grey line) with a moving average (black line) fitted to each graph in Figure 2.

Cis-element selection

We focused on 22 previously reported, putative cis-elements, most of which have not been functionally characterized, and we refer to these from this point forward as cis-elements without ascribing any functional annotation (Table 1). First, putative cis-elements were selected based on their over-representation in *CesA* promoters associated with primary (CRPE17, CRPE 12, CRPE 11, CRPE 10, CRPE8, CRPE6) or secondary (CRPE31, CRPE28, CRPE26, CRPE25) cell wall formation, as identified in a comparative study of *Eucalyptus*, *Arabidopsis* and *Populus CesA* promoters (Creux et al. 2008). A second set of cis-elements was identified by using the PLACE database homology search tool (http://www.dna.affrc.go.jp/PLACE/, Higo et al. 1999). Cis-elements with motif lengths of

greater than five base pairs were selected if they were present in the promoters of all three primary or all three secondary cell wall-related *CesA* genes (Table 1). The xylem-specific promoter element identified by Ko et al (2006) and the tracheary element-specific motif identified by Pyo et al (2007) were also added to the set of cis-elements as they have been suggested to play a role in secondary cell wall formation in *Arabidopsis*.

Cis-element mapping

Pattern Matching in RSA-Tools (http://rsat.scmbb.ulb.ac.be/rsat/; Thomas-Chollier et al. 2008) was used to map occurrences of the selected cis-elements onto the six *CesA* promoter sequences of each of 13 *Eucalyptus* species. Pattern matching and image generation were conducted using the default settings. A random data set was generated using RSA-Tools Random Sequence Generator, which calibrated the sequences on *Arabidopsis* non-coding upstream sequences. An *Arabidopsis*-specific Markov model was used to generate the random sequences. In order to visualise the cis-element maps generated by RSA Tools (Figure 3) in a meaningful way so that patterns of cis-element conservation and variation could be more easily observed we concatenated the promoter sequences from each gene together per species. These sequences were then used to construct a preliminary neighbour joining tree in MEGA (Tamura et al. 2011) and the species order was used to arrange the promoters during cis-element mapping (Figure 3). Promoter sequences containing the mapped cis-elements were divided into discrete 100 bp intervals (i.e. -1 to -100, -101 to -200, etc.), and cis-element occurrences in each section were counted and graphically represented in comparison with the random data set. A two-tailed t-test assuming equal variance was performed, to identify regions that showed significant differences from the random dataset with $\alpha = 0.01$ and $\alpha = 0.001$. Motif logos were generated using the output sequences from RSA-Tools in the online motif logo tool, Weblogo (http://weblogo.berkeley.edu/logo.cgi), with all default settings (Schneider and Stephens 1990; Crooks et al. 2004). The *Eucalyptus*-

specific motif logos were compared to the cis-element consensus logos generated across all three genera (*Arabidopsis*, *Populus* and *Eucalyptus*) using the same tools (Online Resource 4).

Cis-element conservation analysis

Cis-element conservation was estimated by counting the number and type of nucleotide changes that occurred within cis-element occurrences (Online Resource 5 and 6). The cis-element occurrences could be grouped into three categories: conserved cis-element occurrences, where no changes were observed in any of the *Eucalyptus* species analysed; moderately conserved cis-element occurrences, where only a single position in the cis-element sequence was changed in one or more species; and non-conserved cis-element occurrences, where more than one position in the cis-element sequence was changed in one or more of the species analysed. The types of nucleotide changes were also classified as singletons (occurring in only one of the 13 species), or polymorphisms (occurring in two or more species). The different cis-element and mutation counts were entered into Excel (Microsoft Office 2007), where the averages and percentages of the different polymorphism affecting cis-element occurrences were calculated (Online Resource 6).

Results

Sequence divergence and nucleotide diversity of the *Eucalyptus CesA1* gene and promoter at the population, species and genus levels

We compared the nucleotide diversity of the proximal promoter and 5' regions (ATG was the anchor point placed at position 0 bp) of the secondary cell wall-related *Eucalyptus CesA1* gene (Ranik and Myburg 2006), including representative portions of the 5' UTR, first exon and first intron to the orthologous cellulose synthase genes of *Arabidopsis* and *Populus* (Figure 1). Nucleotide diversity in these regions was calculated for three sets of sequences: (1) the genus-level comparison of the *E. urophylla CesA1* gene to its *Arabidopsis* and

Populus orthologs (Online Resource 3), (2) the species-level comparison of the *CesA1* gene sequences from thirteen *Eucalyptus* species (Online Resource 1) and finally, (3) a population-level comparison which included sequences from a population sample of 19 *E. urophylla* trees (Figure 1). As expected, the average nucleotide diversity over the whole 611 bp region was highest for the genus-level comparison ($\pi = 0.461$), which was close to a value expected for unrelated sequences (Figure 1, green trend line). The *Eucalyptus* species- and population-level comparisons exhibited significantly lower average diversity ($\pi = 0.015$ and $\pi = 0.006$, respectively) in the same region. The observed nucleotide diversity at the population level (Figure 1, black trend line) was generally below 1% and the diversity observed in parts of the promoter and first exon was due to single nucleotide polymorphisms (SNPs). The nucleotide diversity of the promoter was higher at the species-level (Figure 1, red trend line), but distinct regions of relative sequence conservation were observed within the promoter region (Figure 1, black arrows). These results demonstrate that at the genus-level, the promoters are too divergent structurally to observe conserved sequence elements by direct sequence comparison. Conversely, the population-specific comparison suggested that the sequences were too similar to identify defined regions of conservation. The species-level comparison, however, did reveal conserved regions in the promoters, which may contain clusters of cis-regulatory elements and the other five *CesA* promoters (*CesA2*, *3*, *4*, *5* and *7*) were consequently investigated at this level.

Isolation and analysis of the *CesA1-5* and *7* promoter regions of 13 *Eucalyptus* species

The upstream regions of six *CesA* genes (*EgCesA1-5*: Ranik and Myburg 2006; *EgCesA7*: Creux et al. 2008) were isolated from one individual of each of 13 *Eucalyptus* species and used to investigate detailed patterns of sequence and spatial conservation of selected, putative cis-elements among the *Eucalyptus* species (sequence alignment, Online Resource 5). Creux et al (2008) showed that *E. grandis CesA4* had an intron in the 5'UTR and this intron was

also present in the CesA4 promoter of the other 12 *Eucalyptus* species investigated here indicating conservation in the 5'UTR in the gene. It is well documented that 5' UTR sequences and the associated introns can play important roles in cis-regulation (Karthikeyan et al. 2009; Livny and Waldor 2009). For this reason, we disregarded promoter/ UTR boundaries and in all instances, the entire upstream regions including the 5' UTR (and intron in the case of *CesA4*) were analysed for regulatory element occurrence and conservation.

Despite several rounds of primer optimisation and design, we were unable to isolate the upstream regions of the *CesA5* gene from *E. camaldulensis, E. tereticornis* and *E. dunnii* presumably due to high sequence divergence in the upstream priming sites. Amplification of the *E. fastigata CesA5* promoter was only achieved when the primers were moved immediately upstream of the original binding sites, which were subsequently found to contain *E. fastigata*-specific sequence polymorphisms (primer sequences, Online Resource 2). As a result, the *CesA5* promoter dataset only contained sequences from ten *Eucalyptus* species. Furthermore, two distinct sequence haplotypes of the *CesA5* promoter were observed among the ten *Eucalyptus* species analysed (Online Resources 4 and 5). Species from the subgenus *Eucalyptus* (*E. fastigata*) and *Symphyomyrtus* section *Latoangulatae* (*E. grandis, E. urophylla* and *E. saligna*) shared a single haplotype. The second haplotype was only observed in species of the section *Maidenaria* (*E. macarthurii, E. globulus maidenii, E. globulus globulus, E. globulus bicostata, E. smithii* and *E. nitens,* Online Resources 4 and 5). PCR amplification with haplotype-specific primers confirmed that the two putative haplotypes do not co-occur within any of the eucalypt species analysed and are therefore not likely to be derived from paralogous sequences in the *Eucalyptus* genome (data not shown).

There are 11 expressed *CesA* genes in *Eucalyptus* (Mizrachi et al. 2010) and to ensure that no duplicate promoters were isolated all of the *CesA* promoter sequences analysed in this study were compared to the *E. grandis* genome sequence (DOE JGI,

http://www.phytozome.net). In all cases the sequences matched a single region directly upstream of the corresponding *CesA* gene, supporting our inference that no paralogous promoters were isolated. Sequence comparison of the two *CesA5* promoter haplotypes with the genome sequence also confirmed that the two haplotypes correspond to the promoter regions of a single *CesA5* gene locus in the *E. grandis* genome (data not shown, Phytozome gene ID: Eucgr.C02801.1).

Species-level sequence diversity in the promoter regions of the *Eucalyptus CesA* genes

The average species-level nucleotide diversity (π) of the six *Eucalyptus CesA* promoter regions varied from $\pi = 0.014$ for *CesA7* to $\pi = 0.068$ for *CesA5* (Table 2). The high species-level nucleotide diversity observed in the *CesA5* promoter regions could be ascribed to the presence of the two distinct haplotypes (Online Resource 5 and 7). In all of the promoters a local decrease in nucleotide diversity across species was observed at the transcriptional start site (Figure 2, shaded boxes). Additionally, in the *CesA1*, *2*, *3*, *4* and *7* promoters there were regions further upstream where local species-level diversity was below $\pi = 0.02$, which is the nucleotide diversity expected for conserved coding regions and could indicate functional constraints within these regions (Figure 2).

Cis-element position and frequency in the promoter regions of the *Eucalyptus CesA* genes

We investigated the positional conservation of cis-elements in the *CesA* promoter regions by mapping occurrences of 22 previously identified putative cis-elements (Table 1) in the promoter sequences of 13 *Eucalyptus* species (Figure 3 and Online Resource 5). Three of the cis-elements (CRPE25, CRPE26 and TERE;Table 1) could not be found in any of the 13 *Eucalyptus CesA* promoter sequences, even when mismatches were allowed, but they were detected in the *Arabidopsis* and *Populus* orthologs (Figure 3). The CRPE11 cis-element could not be identified in any of the sequences analysed and may represent a false positive result

from the previous study. Cis-element occurrence counts revealed regions with a significantly higher/lower number of occurrences in the promoter sequences, while an even cis-element distribution was observed in the random dataset (Figure 4). The occurrences of each cis-element in the *Eucalyptus CesA* promoters were used to generate *Eucalyptus*-specific cis-element consensus sequences which were very similar at the species- and genus-level, even in sites allowing for alternative bases (Online Resource 4).

Mapping the cis-element occurrences to the promoter regions (Figure 3 and Online Resource 5) allowed us to evaluate the cis-element content of conserved promoter regions (Figure 2). In each of the *Eucalyptus* promoters, a set of co-occurring elements (Figure 3, transparent grey blocks) was identified near the previously predicted TSSs (Creux et al. 2008). This set of putative cis-elements comprised multiple occurrences of the CRPE31 (GNGNAGNA, Figure 3: orange) and CTRMCAMV35S (TCTCTCTCT, Figure 3: purple) motifs with the exception of *CesA3* where only CRPE31 occurred approximately 200 bp upstream of the predicted TSS (Figure 3). These were not conserved in the *Arabidopsis* and *Populus* promoters (Figure 3). In *CesA7,* however, the cluster and region of lower species-level nucleotide diversity was located further upstream (-450 to –300) of the TSS and may indicate that the predicted TSS for this gene should be re-evaluated. The TATABOX 5 element (TTATTT, Figure 3: pink), from the PLACE database, was the only TATA-box-like sequence identified in the *Eucalyptus CesA* promoters. In the *CesA1*, *5* and *7* promoters, the putative TATA-box-like motif was found far upstream (400 to 1000 bp) from the TSS (Figure 3).

The cis-element maps also enabled us to observe positional conservation of cis-element combinations. MYB1AT (WAACCA, Figure 3: black) and NODCON1GM (AAAGAT, Figure 3: brown) were found to co-occur in the secondary cell wall-related *CesA* promoters within 200 bp of each other (Figure 3A, B and C). These two putative cis-elements

also appeared to be positionally conserved within these promoters as they were always observed in the region between -200 to -600 bp upstream of the TSS in *CesA1*, *2* and *3* (Figure 3D, E and F).

Cis-element evolution in the promoter regions of the *Eucalyptus CesA* genes

To investigate the potential effect of DNA sequence evolution on cis-element occurrences in *Eucalyptus*, all of the putative cis-element occurrences in the *Eucalyptus CesA* promoters were investigated for sequence conservation. The individual cis-element occurrences were scored as conserved if no polymorphism occurred in the region across all 13 promoter sequences (Figure 5A and Online Resource 6). Putative cis-element occurrences that had a single nucleotide polymorphism (present in two or more cloned sequences) relative to the *Eucalyptus* consensus sequence (Online Resource 4) and those that had two or more nucleotide changes away from the consensus sequence were counted separately (Figure 5A and Online Resource 6). Overall, only 29% of the cis-element occurrences investigated were fully conserved in the 13 *Eucalyptus* promoter sequences. The *Eucalyptus CesA1* promoters had the highest number of fully conserved cis-element occurrences (59%), while *CesA5* had the lowest at 13% (Figure 5A).

We identified 89 instances (42% of all occurrences) where a putative cis-element occurrence was present or absent in all but one of the 13 promoters (Online Resource 6). These instances were classified as singleton (gain or loss) occurrences and 70% of these resulted from single nucleotide changes, while 30% were due to indels (Figure 5B). Some of the single nucleotide changes could have resulted from cloned PCR errors, however the observed frequency of singletons was approximately ten times higher than would be expected from polymerase induced errors (average $1x10^{-4}$ per base pair; Keohavong and Thilly 1989; Ling et al. 1991). We investigated the frequency of singleton mutations that resulted in a change towards the consensus sequence (i.e. a change that rendered the cis-element

identifiable by the software in that promoter) or away from the consensus sequence (i.e. a change rendering the consensus sequence unidentifiable by the software). The majority of sequence changes (72% of the SNPs and 71% of the indels, Figure 5B) resulted in a change towards the cis-element consensus sequence (Online Resource 4) and therefore may indicate a gain (or maintenance) of the cis-element occurrence in that species. However, to accurately investigate this, more individuals should be sequenced for each species and experimentally verified *Eucalyptus* cis-elements, which are not yet available, should be used.

Discussion

CesA gene family members have conserved roles in the deposition of primary and secondary cell walls in all seed plants studied to date (Burn et al. 2002; Burton et al. 2004; Hamann et al. 2004; Ranik and Myburg 2006). This suggests that the major clades of the gene family differentiated early during Spermatophyte evolution (Sarkar et al. 2009; Yin et al. 2009; Popper et al. 2011). A distinctive set of expression patterns characterize the major clades with *CesA* genes involved in primary and secondary cell wall deposition exhibiting unique, developmentally regulated expression profiles (Hamann et al. 2004). The conserved nature of these genes and their highly coordinated, but differential, expression patterns point to the action of a conserved network of cis- and trans-regulatory factors in plants (Demura and Fukuda 2007; Zhong et al. 2010). The aim of this study was to characterize the architecture and diversity of cis-elements in six *CesA* promoters across *Eucalyptus* tree species in terms of sequence and cis-element conservation.

Nucleotide diversity levels are generally higher in promoter regions than in other genic regions, most likely due to lower overall functional constraints on promoter regions (Nei 2007). That said, the maintenance of functional cis-elements in promoter regions and the modular nature of transcription factor binding suggests that there should be localised regions

that are more conserved than the rest of the promoter and contain clusters of cis-elements (Maniatis et al. 1987; Ho et al. 2009). To determine whether this is the case for *CesA* genes, we compared nucleotide diversity levels in different *CesA1* gene regions at the population, species and genus levels. While the nucleotide diversity in the genus and population levels were at the highest and lowest extremes of the scale respectively, the diversity at the species level showed a large range of values from 1.4% to 6.8% (Figure 1). At the species level the nucleotide diversity plot profile revealed that the higher nucleotide diversity in the upstream regions was interrupted by areas with distinctly lower nucleotide diversity (Figure 1 - red line and arrows). In a similar study on the *Drosophila* bithorax complex, Ho et al (2009) also identified regions of conservation across orthologous promoters in different *Drosophila* species and could suggest that short conserved sequences are a feature of related eukaryotic promoters.

The conservation of promoter sequences has been documented for a number of different plant species, but sequence conservation in promoter regions may not always indicate cis-element conservation (Reineke et al. 2011). In light of this, we evaluated the occurrences of previously identified cis-elements (Table 1) in the cloned *Eucalyptus* promoter sequences. We found that 76% of the cis-element occurrences were either fully conserved (29%) in the 13 *Eucalyptus* promoters or only varied from the consensus sequence by a single nucleotide change (47%, Figure 5A and Online Resource 6). In this study the cis-element occurrences which are affected by one nucleotide change are likely also to be conserved elements because cis-element consensus sequences often contain some ambiguous bases and transcription factors can still bind to these variable sites.. The cis-element occurrences with a single change could also be more conserved than estimated here since the changes may be due to allelic variation within the different populations, but this variation

would have to be further investigated in these different populations when a core set of experimentally tested cis-elements are available for *Eucalyptus*.

Identifying conserved cis-elements or regions in the promoter does not fully describe the spatial arrangement of these elements, which is important because cis-elements are often found clustered together, rather than evenly distributed across the length of the promoter (Maniatis et al. 1987; Hansen et al. 2010). We found that there was strong clustering of cis-elements at particular intervals in the promoters when compared to neighbouring intervals or to a random dataset (Figure 4). A significant cluster of cis-elements was observed at the position of the TSS in a number of the *CesA* promoters (Figure 4: *CesA1*, *4*, *5* and *7*), and the position of this cis-element cluster (TSS-associated cluster) also coincided with a highly conserved promoter region (Figure 2; transparent grey blocks). In a genome-wide comparative study of *Populus* and *Arabidopsis* a number of conserved cis-element clusters were also identified in the promoters of genes associated with cellulose deposition (Ding et al. 2012) and this could indicate functional constraints acting on the particular *CesA* promoter regions.

The TSS-associated cluster of cis-elements identified in five of the six *Eucalyptus CesA* promoters (except *CesA3*) contained multiple occurrences of two cis-elements CRPE31 and CTRMCAMV35S (Figure 3: purple). CRPE31 (GNGNAGNG, Figure 3: orange) was the most abundant cis-element detected in the *Eucalyptus* promoters as well as in a previous cis-element study of *CesA* promoters (Creux et al. 2008). The reverse compliment sequence of this element (CNCTNCNC) could be an initiator element similar to the elements described in *Arabidopsis* (Bernard et al. 2010). The initiator element has been shown to co-occur with the TSS in the promoters of different plant species (Yamamoto et al. 2007) and this may suggest a function for these putative CRPE31 occurrences in the *Eucalyptus* promoters.

The other main element in the TSS-associated cluster was the CTRMCAMV35S element (TCTCTCTCT, Figure 3: purple). In the PLACE database, this element is listed as an enhancer found in the commonly used Cauliflower Mosaic Virus (CAMV) 35S viral promoter (Pauli et al. 2004). This element is highly over-represented in the *CesA* promoters in the form of TC-repeats (Online Resource 5). Multiple copies of the CTRMCAMV35S element could enhance the expression of the *CesA* genes which would be of interest because these are some of the most highly expressed *Eucalyptus* genes (Mizrachi et al. 2010). Alternatively, the CTRMCAMV35S repeats could represent plant-specific regulatory regions known as Y-patches (CT or TC repeats) which have been associated with the TSSs of other plant promoters in genome-wide studies of *Arabidopsis* and rice (Molina and Grotewold 2005; Yamamoto et al. 2007). The putative CRPE31 and CTRMCAMV35S element occurrences in this conserved TSS-associated cluster may play a role in the initiation of transcription of a number of the *Eucalyptus CesA* genes.

The suggestion that transcriptional initiation of the *Eucalyptus CesA* genes could be reliant on the presence of putative initiator (CRPE31) and Y-patch (CTRMCAMV35S) elements is further bolstered by the lack of a TATA-box in many of the *Eucalyptus CesA* promoters (Figure 3). The only over-represented element resembling a TATA-box (PLACE ID: TATABOX5; TTATTT) was found in a subset of the *CesA4* promoter sequences in the correct location (Figure 3: pink). This TATA-box-like element was also identified in the *Eucalyptus CesA1*, *4*, *5* and *7* promoters, but its position in each instance was ≥ 600 bp upstream of the TSS (Figure 3, pink) and did not coincide with any major decrease in species-level nucleotide diversity. This suggests that these sequences do not function as TATA-boxes, because in the original description of this element it was functional when located 50 bp upstream of the TSS (Tjaden et al. 1995). Additional support for this TSS-associated cluster comes from our previous study which reported some of the *Eucalyptus*

grandis CesA promoters, which contain the TSS-associated cluster, to be TATA-less. However, the promoter fragments were still able to drive tissue-specific expression of the GUS reporter gene in *Arabidopsis* and had experimentally verified TSSs (Creux et al. 2008).

Further upstream of the TSSs, other putative cis-elements could be observed as singletons, pairs or clusters in the *CesA* promoters (Figure 3). One example of this is the co-occurrence of the NODCON1GM (AAAGAT, Figure 3: dark brown) and MYB1AT (WAACCA, Figure 3: black) elements in the *CesA* promoters of the genes associated with secondary cell wall deposition. In *CesA1*, *2* and *3*, the elements co-occurred in the region between -200 and -600 bp, and this positional conservation was not observed in the promoters of the *CesA* genes associated with primary cell wall formation (Figure 3). This suggests that NODCON1GM and MYB1AT may play a role in secondary cell wall-specific expression of these genes.

The putative MYB1AT occurrences in *CesA1*, *2* and *3* were of special interest for a number of reasons. Yazaki et al. (2003) found the MYB1AT element to be involved in *Arabidopsis* dehydration stress response while investigating gibberellic acid (GA) and abscisic acid (ABA) responses in rice. GA has also been found to play a role in cambial cell differentiation and xylem development (Love et al. 2009) pointing to the co-regulation of genes involved in xylogenesis. This putative MYB-like element may also be important as a number of MYB transcription factors have been identified as key members of the transcriptional network regulating the secondary cell wall formation (Goicoechea et al. 2005; Demura and Fukuda 2007; Legay et al. 2010; Zhong et al. 2010).

While many of the putative cis-element occurrences investigated in this study were conserved in the promoters of the different *Eucalyptus* species, a small number of occurrences varied among species (Figure 5B). We counted 89 singleton changes (i.e. loss or gain of an occurrence in only one of the 13 sequences) in the *CesA* promoters and 66% of the

sequence polymorphisms changed the promoter sequence towards a known cis-element consensus sequence (Online Resource 6). One hypothesis for this cis-element variation is known as cis-element buffering, which ensures the maintenance of a particular cis-element sequence in faster evolving sequences such as promoters (Tanay et al. 2005). It proposes that the promoter sequence may change to abolish a particular cis-element occurrence, but a second mutation in that promoter maintains the binding site of a particular transcription factor. This mechanism of cis-element maintenance has been observed in the cis-regulatory modules of a number of *Drosophila* genes (Hare et al. 2008; Ho et al. 2009) and could be a mechanism of maintenance in the *CesA* gene promoters, but this can only be confirmed by further investigation on both the population-wide and genome-wide level with experimentally verified cis-elements from *Eucalyptus*.

An interesting observation form this study was that, while most of the *Eucalyptus CesA* promoters were similar in terms of GC content and number of indels, the *CesA5* promoters were distinct. The *CesA5* promoter dataset had the lowest GC content (36%) and more than double the number of indels than the other promoters (Table 2). *CesA5* also presented with the lowest number of conserved cis-element occurrences (Figure 5A). Sequence analysis revealed at least two *CesA5* promoter haplotypes in the *Eucalyptus* dataset and these were congruent with the different sections of *Symphyomyrtus* in which they occurred (Online Resources 1 and 5). The promoter regions of species belonging to the section *Exsertaria* (*E. camaldulensis* and *E. tereticornis*) could not be amplified and might represent yet another haplotype. It appears that the distal part of the *CesA5* promoter is undergoing rapid divergence, and the pattern of divergence is in keeping with the phylogeny of the *Eucalyptus* species. Since haplotype 1 is shared among representatives of both subgenera, it is likely to be ancestral and the alignment of the *CesA5* promoter sequences (Online Resources 4 and 5) suggests that multiple insertional events have occurred. The high

sequence divergence observed in the distal region (upstream from -390 bp) suggests that the proximal region of the *CesA5* promoter may harbour most of the essential cis-elements for this gene.

Conclusions

This is one of only a few studies that have investigated sequence diversity in the promoters of a plant gene family (Koch et al. 2001; de Meaux et al. 2006; Tanaka et al. 2009; Zhao et al. 2009). We identified regions within the *CesA* promoters that are conserved across *Eucalyptus* species and coincided with the putative occurrences of cis-elements, suggesting that they have important biological functions. Overall, we found that 29% of the investigated cis-element occurrences were fully conserved in *Eucalyptus CesA* genes. Only 30% of the singleton changes were away from the consensus sequence (Figure 5B), suggesting there are functional constraints on some sequences within the promoter regions. The *CesA* promoters of *Eucalyptus* appear to be TATA-less and a highly conserved region in these promoters was identified in the vicinity of the TSSs, suggesting that the basal transcriptional machinery for this *Eucalyptus* gene family relies on other basal cis-elements such as a putative initiator element and Y-patch to initiate transcription. Conserved cis-elements were also found in the promoters of the *CesA* genes associated with secondary cell wall formation. These were not present in the primary cell wall-associated *CesA* promoters. The study provides insight into the diversity and evolution of cis-regulatory sequences underlying the unique expression profiles of this important plant gene family and will lay the foundation for future studies of the function of these promoter regions and for comparative genomic analysis of promoter elements across multiple *Eucalyptus* genomes.

Acknowledgements

The authors would like to acknowledge Alisa Postma for her contribution to the preliminary study leading to this research and Dr. Albe van der Merwe and Dr. Christine Martiz-Olivier for comments and guidance during the preparation of the manuscript. This work was supported with funding provided by Mondi and Sappi, through the Forest Molecular Genetics (FMG) Programme, the Technology and Human Resources for Industry Programme (THRIP) and the National Research Foundation of South Africa (NRF).

Tables

Table 1. Details of 22 cis-regulatory elements selected from literature and PLACE database scans and used for DNA pattern matching

Source [a]	Motif Identity [b]	Motif Sequence [c]	PLACE Annotation [d]
Primary cell wall-associated motifs[e]			
Creux et al. (2008)	CRPE 17	GTCKGT	Unknown
Creux et al. (2008)	CRPE12	ATNWATTA	Phosphate response domain
Creux et al. (2008)	CRPE11	GACNGTSNGTGGGC	Stem enhancer element
Creux et al. (2008)	CRPE10	CCNCMCCC	Vascular-specific expression
Creux et al. (2008)	CRPE8	GGNGGTGG	Anthocyanin regulatory element
Creux et al. (2008)	CRPE6	NMTTCTGTC	Iron deficiency responsive element
Place DB	CAREOSREP1	CAACTC	Gibberellin up-regulated proteinase expression
Place DB	DRE2COREZMRAB17	ACCGAC	Drought-responsive element
Place DB	RBCSCONSENSUS	AATCCAA	Light-regulated expression
Place DB	TATABOX5	TTATTT	Functional TATA element
Place DB	SEF4MOTIFGM7S	RTTTTTR	Beta-conglycinin enhancer
Secondary cell wall-associated motifs[e]			
Creux et al. (2008)	CRPE31	GNGNAGNG	Unknown
Creux et al. (2008)	CRPE28	NNGCATGC	Iron deficiency response element
Creux et al. (2008)	CRPE26	TCCTGCYG	Unknown
Creux et al. (2008)	CRPE25	RCYSTGCCC	Phloem-specific expression
Place DB	CTRMCAMV35S	TCTCTCTCT	Enhancer of gene expression
Place DB	REALPHALGLHCB21	AACCAA	Phytochrome regulatory elements
Place DB	PYRIMIDINEBOXOSRAMY1A	CCTTTT	Pyrimidine box involved in sugar repression
Place DB	NODCON1GM	AAAGAT	Organ-specific element
Place DB	MYB1AT	WAACCA	Activation draught and ABA-induced expression
Pyo et al. (2007)	TERE	CTTNAAAGCNA	Tracheary element-specific expression
Ko et al.(2006)	XYLAT	ACAAAGAA	Xylem-specific

[a] Original source of the cis-element

[b] Published name or identity of the cis-element

[c] Published consensus sequences for the cis-element motifs with ambiguous bases represented as IUPAC codes where W = A/T, M = A/C, R = A/G, K = T/G, S = G/C, Y = C/T and N represents any of the four bases.

[d] Putative function of the cis-elements as reported in literature or the PLACE database.

[e] The cis-element motifs were assigned as primary or secondary cell wall-associated based on the study in which they were first identified and/ or the description on the PLACE database.

Table 2. Species-level nucleotide diversity in the promoter regions of six cellulose synthase (*CesA*) genes cloned from 13 *Eucalyptus* tree species

	CesA1	***CesA2***	***CesA3***	***CesA4***	***CesA5***	***CesA7***
Number of species analysed	13	13	13	13	10[b]	13
Length of aligned sequence (including gaps)	1132	1284	1317	1286	1970	863
G+C content (%)	43	47	45	47	36	53
Total number of sites (excluding gaps)	989	1048	1238	1168	1260	730
Number of polymorphic sites	66	134	93	105	253	51
Total number of singleton sites	72	138	95	110	261	52
Nucleotide diversity (π)	0.018	0.029	0.018	0.021	0.068	0.014
Nucleotide diversity (θ_W)	0.023	0.042	0.025	0.030	0.073	0.023
Total number of insertions and deletions (indels)[a]	4	6	3	6	15	9

[a] Including repeat regions and indels of varying lengths occurring in more than one species analysed.

[b] The *CesA5* promoter region could only be isolated from 10 species.

Figures

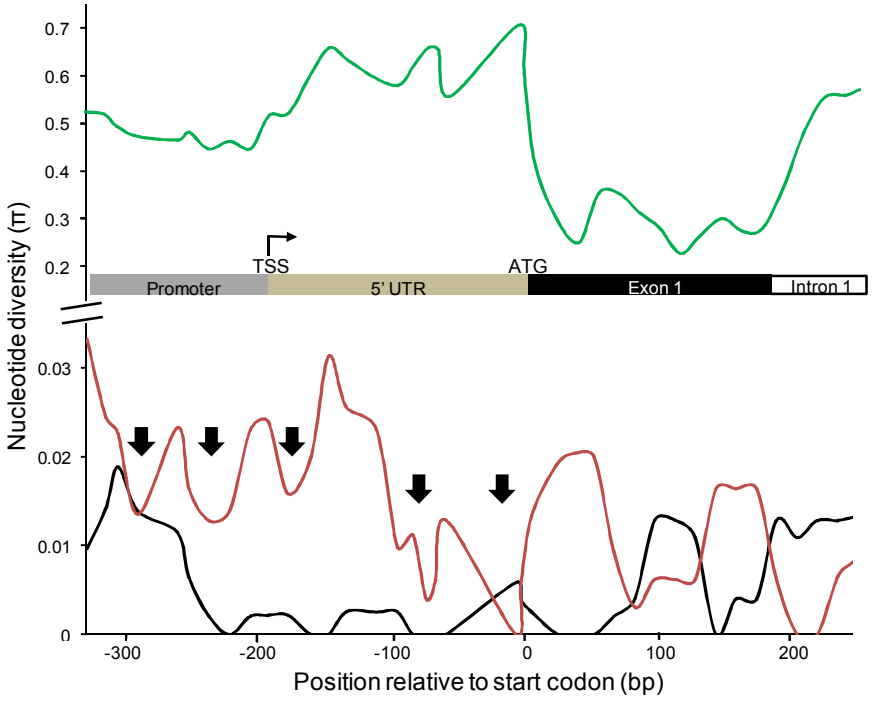


Figure 1. Nucleotide diversity (π) in the proximal promoter and gene regions surrounding the translational start site of the *Eucalyptus CesA1* gene (orthologous to *CesA8* in *Arabidopsis* and *Populus*) at the genus, species, and population levels (Accession numbers: JN573752 - JN573783). The gene regions are depicted by the horizontal bar at the centre of the graph. Nucleotide diversity among *Eucalyptus urophylla CesA1* and its *Arabidopsis thaliana* (*AtCesA8*: At4g18780, TAIR9 - www.arabidopsis.org) and *Populus trichocarpa* (PtiCesA8-A: Pti235238; Kumar et al. 2009) orthologs is shown in green. Nucleotide diversity in the corresponding regions of the *CesA1* gene from thirteen *Eucalyptus* species is shown in red, while nucleotide diversity in the *CesA1* gene from a population of *E. urophylla* trees is shown in black. In each case, nucleotide diversity is represented by the average of a moving window of 50 bp, calculated in DnaSP (DNA Sequence Polymorphism, version 4.50.3, Rozas et al. 2003). The black arrows indicate promoter and 5' UTR regions with lower nucleotide diversity for the species comparison (red line) indicating putatively informative regions of conservation. TSS - transcriptional start site, ATG - translational start site.

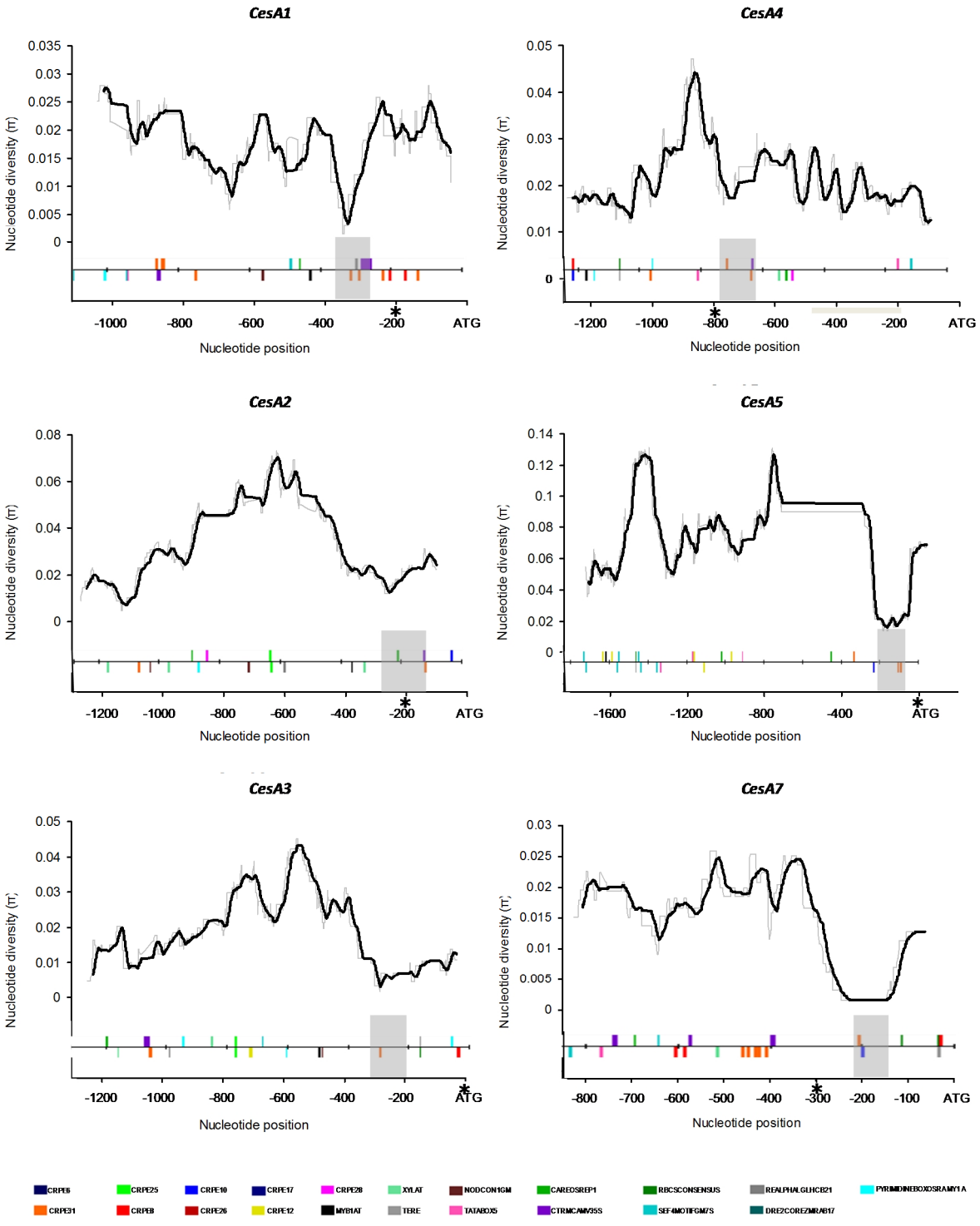
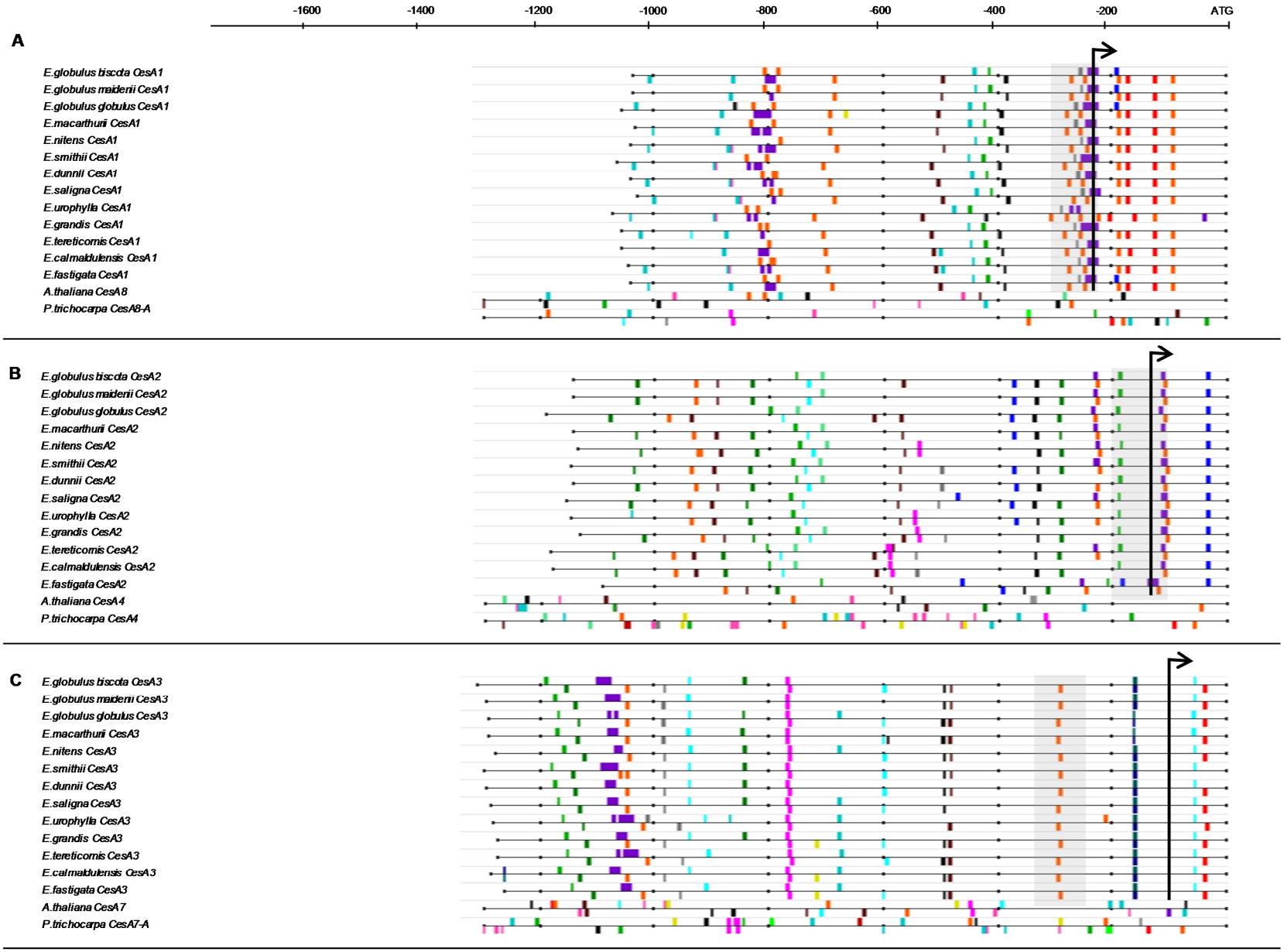


Figure 2. Nucleotide diversity profiles of the promoter regions of six cellulose synthase (*CesA*) genes in 13 *Eucalyptus* species and mapped positions of putative cis-elements in these regions. Nucleotide diversity (π) was measured per site in a sliding window of 100 bp (grey line) with a moving average (black line) fitted to each graph as calculated in DnaSP (DNA Sequence Polymorphism, version 4.50.3, Rozas et al. 2003). Nucleotide position is indicated relative to the start of translation (ATG, +1). The putative transcriptional start site (TSS) is indicated by an asterisk in each graph. A conserved intron present in the 5' UTR of *CesA4* is indicated with a grey bar. The line and coloured blocks at the bottom of each graph show the position of the mapped cis-element occurrences in the *E. grandis* reference sequence. A cis-element colour key is at the bottom of the figure. The transparent grey blocks represent the position of the TSS-associated cis-elements associated with regions of lower nucleotide diversity.


-1600
-1400
-1200
-1000
-800
-600
-400
-200
ATG
A
E.globulus tiscota CesA1
E.globulus maidenii CesA1
E.globulus globulus CesA1
E.macarthurii CesA1
E.nitens CesA1
E.smithii CesA1
E.dunnii CesA1
E.saligna CesA1
E.urophylla CesA1
E.grandis CesA1
E.tereticornis CesA1
E.calmaldulensis CesA1
E.fastigata CesA1
A.thaliana CesA8
P.trichocarpa CesA8-A
B
E.globulus tiscota CesA2
E.globulus maidenii CesA2
E.globulus globulus CesA2
E.macarthurii CesA2
E.nitens CesA2
E.smithii CesA2
E.dunnii CesA2
E.saligna CesA2
E.urophylla CesA2
E.grandis CesA2
E.tereticornis CesA2
E.calmaldulensis CesA2
E.fastigata CesA2
A.thaliana CesA4
P.trichocarpa CesA4
C
E.globulus tiscota CesA3
E.globulus maidenii CesA3
E.globulus globulus CesA3
E.macarthurii CesA3
E.nitens CesA3
E.smithii CesA3
E.dunnii CesA3
E.saligna CesA3
E.urophylla CesA3
E.grandis CesA3
E.tereticornis CesA3
E.calmaldulensis CesA3
E.fastigata CesA3
A.thaliana CesA7
P.trichocarpa CesA7-A
Cis-element key:
CRPE6
CRPE25
CRPE10
CRPE17
CRPE28
XYLAT
NODCON1GM
CAREOSREP1
RBCSCONSENSUS
REALPHALGLHCB21
PYRIMIDINEBOXOSRAMY1A
CRPE31
CRPE8
CRPE26
CRPE12
MYB1AT
TERE
TATABOX5
CTRMCAMV35S
SEF4MOTIFGM7S
DRE2COREZMRAB17

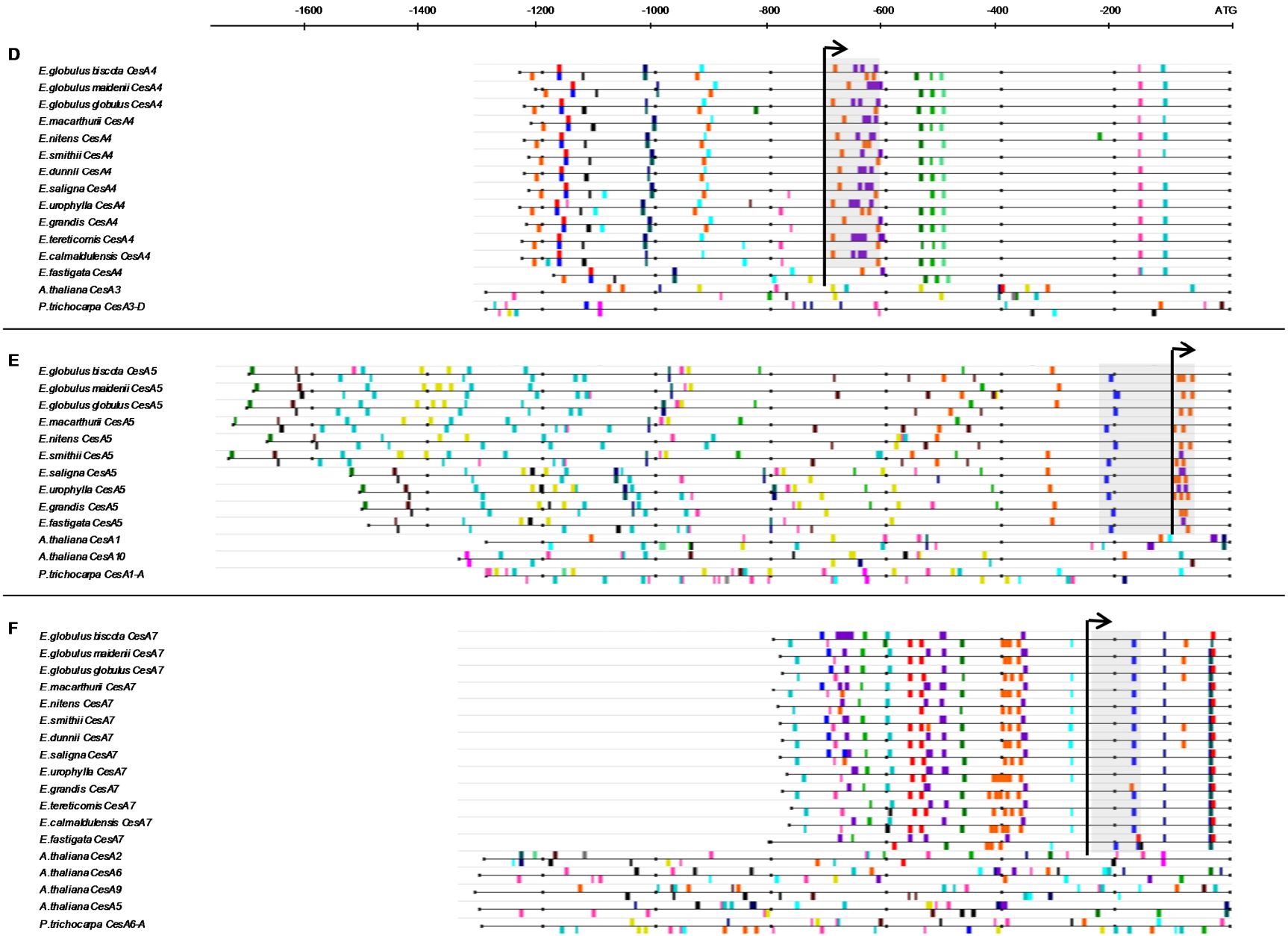

-1600
-1400
-1200
-1000
-800
-600
-400
-200
ATG
D
E.globulus tiscota CesA4
E.globulus maidenii CesA4
E.globulus globulus CesA4
E.macarthurii CesA4
E.nitens CesA4
E.smithii CesA4
E.dunnii CesA4
E.saligna CesA4
E.urophylla CesA4
E.grandis CesA4
E.tereticornis CesA4
E.calmaldulensis CesA4
E.fastigata CesA4
A.thaliana CesA3
P.trichocarpa CesA3-D
E
E.globulus tiscota CesA5
E.globulus maidenii CesA5
E.globulus globulus CesA5
E.macarthurii CesA5
E.nitens CesA5
E.smithii CesA5
E.saligna CesA5
E.urophylla CesA5
E.grandis CesA5
E.fastigata CesA5
A.thaliana CesA1
A.thaliana CesA10
P.trichocarpa CesA1-A
F
E.globulus tiscota CesA7
E.globulus maidenii CesA7
E.globulus globulus CesA7
E.macarthurii CesA7
E.nitens CesA7
E.smithii CesA7
E.dunnii CesA7
E.saligna CesA7
E.urophylla CesA7
E.grandis CesA7
E.tereticornis CesA7
E.calmaldulensis CesA7
E.fastigata CesA7
A.thaliana CesA2
A.thaliana CesA6
A.thaliana CesA9
A.thaliana CesA5
P.trichocarpa CesA6-A
Cis-element key:
CRPE6
CRPE25
CRPE10
CRPE17
CRPE28
XYLAT
NODCON1GM
CARECOSREP1
RBCSCONSENSUS
REALPHALGLHCB21
PYRIMIDINEBOXOSRAMY1A
CRPE31
CRPE8
CRPE26
CRPE12
MYB1AT
TERE
TATABOX5
CTRMCAMV35S
SEF4MOTIFGM7S
DRE2COREZMRAB17

Figure 3. Occurrences of 21 putative cis-regulatory elements mapped in the promoters of six orthologous groups of *CesA* genes in 13 *Eucalyptus* species, *Arabidopsis thaliana* and *Populus trichocarpa*. The size of each promoter region and relative positions of mapped cis-elements in relation to the start codon (ATG) of each gene can be read from the ruler at the top. A colour key of cis-elements is given at the bottom of the image. The left hand margin of the figure shows the name and species of each promoter (grouped A-F). The horizontal black lines in each block represent the promoter sequences isolated for each species and the coloured squares show the position of the mapped cis-element occurrences. Squares above each line indicate cis-elements found in the sense orientation, while coloured squares below the lines indicate cis-element occurences found on the opposite strand. The predicted transcriptional start sites (TSSs) of the *Eucalyptus* promoters are indicated by the tailed arrow in each promoter set. The transparent grey blocks show putative cis-regulatory modules that coincide with the TSS positions.

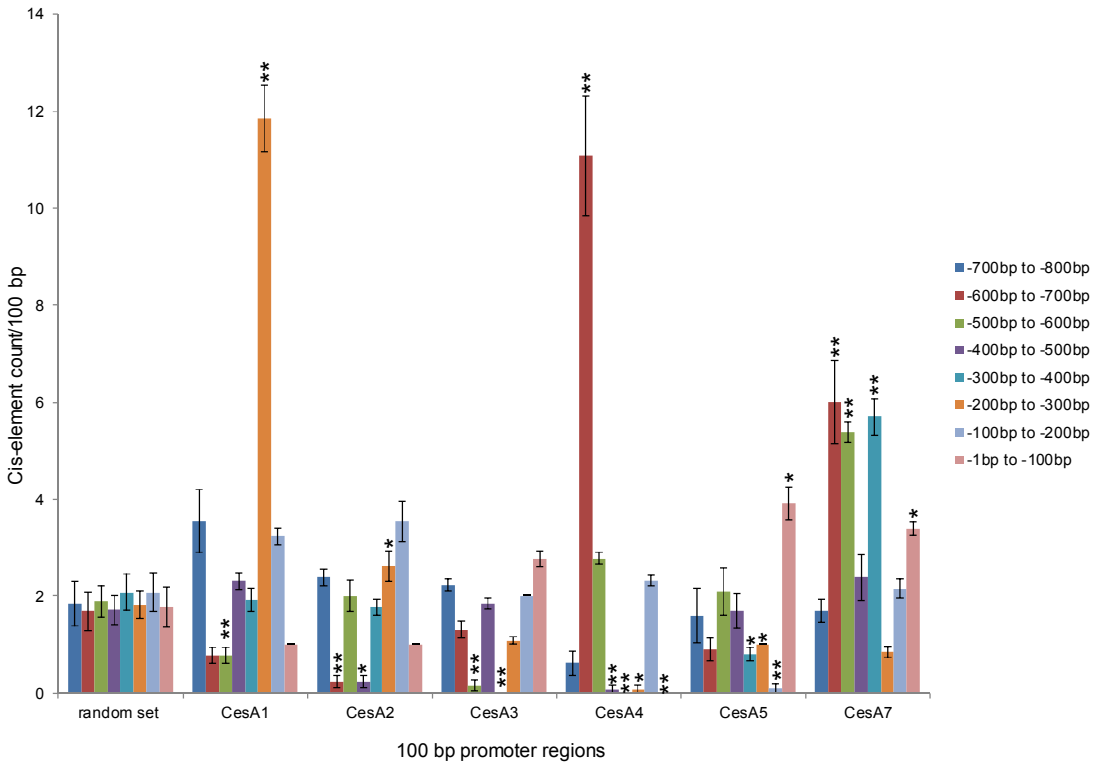


Figure 4. Frequency of cis-element occurrences along the length of the six *CesA* promoter regions averaged across the 13 *Eucalyptus* species in 100 bp intervals, compared to a randomly generated sequence dataset. The random set reports the frequency of occurrences of 21 selected cis-elements in 100 bp intervals along the length of a randomly generated dataset of non-coding upstream sequences (800 bp of upstream from the ATG generated by the RSA-Tools random sequence generator, http://rsat.ulb.ac.be/rsat/). The frequency of cis-element occurrences in each of the *CesA* promoter sets (*CesA1-5* and *7*) are also indicated. The y-axis gives the number of cis-elements in each 100 bp interval and the x-axis represents the *CesA* promoter regions being analysed. The error bars show the standard error of the frequency measurement where $n = 13$ (except for *CesA5* where $n = 10$) and * and ** indicate significant differences from the random dataset ($p = 0.01$ and 0.001 respectively; two tailed t-test assuming equal variance).

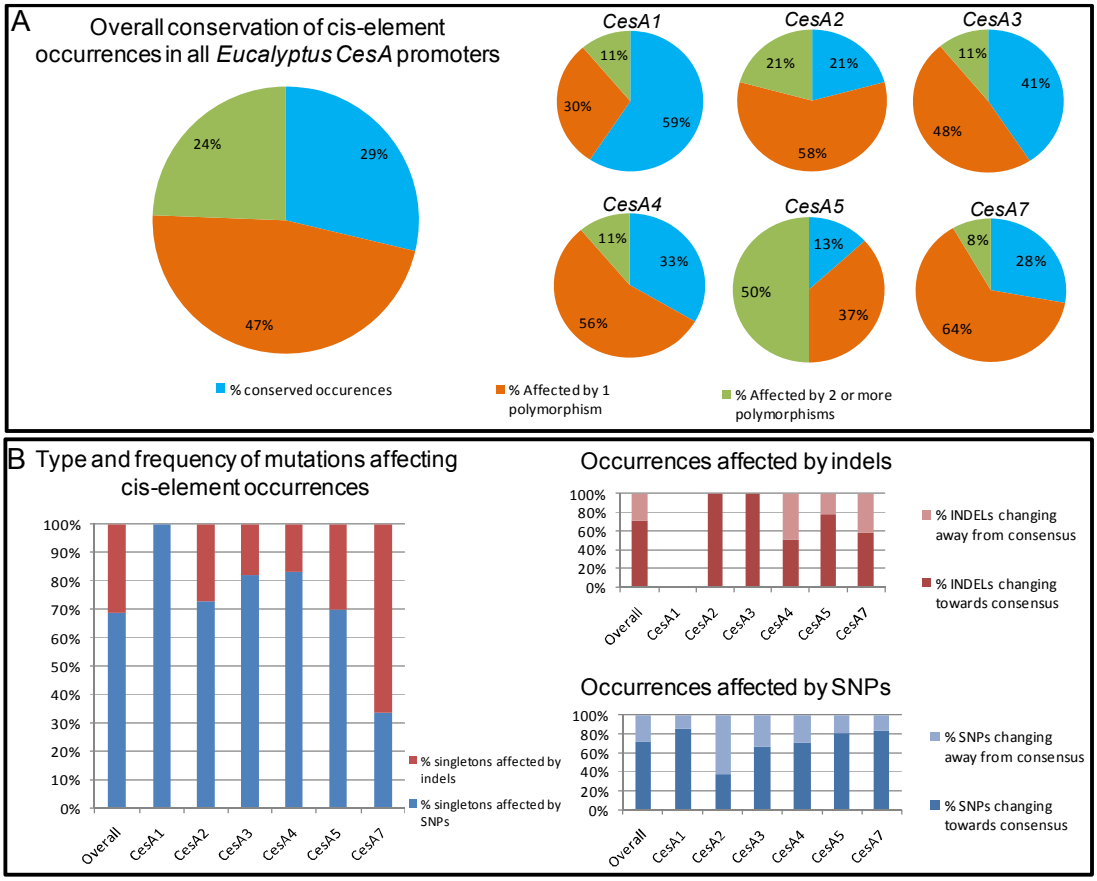


Figure 5. Evaluation of the types and frequency of polymorphisms observed within cis-element occurrences in the *CesA* promoter regions of 13 *Eucalyptus* tree species. A) Conservation of the cis-element consensus sequences across the *Eucalyptus* species. The cis-element occurrences were classified into three categories; conserved occurrences, which had no polymorphisms across all 13 species; occurrences affected by one polymorphism and occurrences affected by two or more polymorphisms (singleton changes were not counted as polymorphisms). The large pie chart on the left depicts the conservation in all promoters across all six genes, while the six smaller pie charts on the right show the cis-element sequence conservation within the promoter of each gene. B) The frequency of single nucleotide polymorphisms and indels in the promoter regions of 13 *Eucalyptus CesA* promoters. For this analysis, polymorphisms in cis-elements that only affect a single promoter of the 13 species (also refered to as singleton gains or losses) were investigated. The graph on the left represents the percentage of cis-element occurrences affected by SNPs and indels. The smaller red bar graph (top right corner) shows cis-element occurrences where an indel changed the sequence towards or away from the consensus sequence. Similarly the blue bar graph (bottom right) indicates cases where SNPs changed the cis-element sequence towards or away from the consensus sequence.

References

Aspeborg H, Schrader J, Coutinho PM, Stam M, Kallas A, Djerbi S, Nilsson P, Denman S, Amini B, Sterky F, Master E, Sandberg G, Mellerowicz E, Sundberg B, Henrissat B, Teeri TT (2005) Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid aspen. Plant Physiol 137: 983-997

Bernard V, Brunaud V, Lecharny A (2010) TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. BMC Genomics 11: 166

Blanchette M, Schwikowski B, Tompa M (2002) Algorithms for phylogenetic footprinting. J Comput Biol 9: 211-223

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

Blanchette M, Tompa M (2003) FootPrinter: a program designed for phylogenetic footprinting. Nucl. Acids Res. 31: 3840-3842

Brooker MIH (2000) A new classification of the genus *Eucalyptus* L'Her. (Myrtaceae). Aust Syst Bot 13: 79-148

Burn JE, Hocart CH, Birch RJ, Cork AC, Williamson RE (2002) Functional analysis of the cellulose synthase genes *CesA1*, *CesA2*, and *CesA3* in *Arabidopsis*. Plant Physiol 129: 797-807

Burton RA, Shirley NJ, King BJ, Harvey AJ, Fincher GB (2004) The *CesA* gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. Plant Physiol 134: 224-236

Carmack CS, McCue L, Newberg L, Lawrence C (2007) PhyloScan: identification of transcription factor binding sites using cross-species evidence. Algorithm Mol Biol 2: 1

Creux NM, Ranik M, Berger DK, Myburg AA (2008) Comparative analysis of orthologous cellulose synthase promoters from *Arabidopsis*, *Populus* and *Eucalyptus*: evidence of conserved regulatory elements in angiosperms. New Phytol 179: 722-737

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188- 1190

Das M, Dai HK (2007) A survey of DNA motif finding algorithms. BMC Bioinformatics 8: S21

de Meaux J, Pop A, Mitchell-Olds T (2006) Cis-regulatory evolution of chalcone-synthase expression in the genus *Arabidopsis*. Genetics 174: 2181-2202

Demura T, Fukuda H (2007) Transcriptional regulation in wood formation. Trends Plant Sci 12: 64-70

Desprez T, Juraniec M, Crowell EF, Jouy H, Pochylova Z, Parcy F, Hofte H, Gonneau M, Vernhettes S (2007) Organization of cellulose synthase complexes involved in primary cell wall synthesis in *Arabidopsis thaliana*. Proc Nat Acad Sci 104: 15572-15577

Ding J, Hu H, Li X (2012) Thousands of cis-regulatory sequence combinations are shared by *Arabidopsis* and poplar. Plant Physiol 158: 145-155

Eldridge K, Davidson J, Harwood C, Wyk GV (1994) Eucalypt domestication and breeding Oxford University Press, New york

Fang F, Blanchette M (2006) FootPrinter3: phylogenetic footprinting in partially alignable sequences. Nucleic Acids Res 34: W617-620

Farnham PJ (2009) Insights from genomic profiling of transcription factors. Nat Rev Genet 10: 605-616

Freeling M, Subramaniam S (2009) Conserved noncoding sequences (CNSs) in higher plants. Curr Opin Plant Biol 12: 126-132

Goicoechea M, Lacombe E, Legay S, Mihaljevic S, Rech P, Jauneau A, Lapierre C, Pollet B, Verhaegen D, Chaubet-Gigot N, Grima-Pettenati J (2005) *EgMYB2*, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis. Plant J 43: 553-567
Gorshkova T, Brutch N, Chabbert B, Deyholos M, Hayashi T, Lev-Yadun S, Mellerowicz EJ, Morvan C, Neutelings G, Pilate G (2012) Plant fiber formation: state of the art, recent and expected progress, and open questions. CRC Cr Rev Plant Sci 31: 201-228
Grattapaglia D, Plomion C, Kirst M, Sederoff RR (2009) Genomics of growth traits in forest trees. Curr Opin Plant Biol 12: 148-156
Hall A (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser: 95-98
Hamann T, Osborne E, Youngs HL, Misson J, Nussaume L, Somerville C (2004) Global expression analysis of *CESA* and *CSL* genes in *Arabidopsis*. Cellulose 11: 279-286
Hansen L, MariÃ±o-RamÃrez L, Landsman D (2010) Many sequence-specific chromatin modifying protein-binding motifs show strong positional preferences for potential regulatory regions in the *Saccharomyces cerevisiae* genome. Nucleic Acids Res 38: 1772-1779
Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid *even-skipped* enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genet 4: e1000106
Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database. Nucleic Acids Res 27: 297-300
Ho MCW, Johnsen H, Goetz SE, Schiller BJ, Bae E, Tran DA, Shur AS, Allen JM, Rau C, Bender W, Fisher WW, Celniker SE, Drewell RA (2009) Functional evolution of *cis*-regulatory modules at a homeotic gene in *Drosophila*. PLoS Genet 5: e1000709
Hu R, Qi G, Kong Y, Kong D, Gao Q, Zhou G (2010) Comprehensive analysis of NAC domain transcription factor gene family in *Populus trichocarpa*. Bmc Plant Biology 10: 145
Karthikeyan AS, Ballachanda DN, Raghothama KG (2009) Promoter deletion analysis elucidates the role of *cis* elements and 5'UTR intron in spatiotemporal regulation of *AtPht1;4* expression in *Arabidopsis*. Plant Physiol 136: 10-18
Keohavong P, Thilly WG (1989) Fidelity of DNA polymerases in DNA amplification. Proc Nat Acad Sci 86: 9253-9257
Kim HD, Shay T, O'Shea EK, Regev A (2009) Transcriptional regulatory circuits: predicting numbers from alphabets. Science 325: 429- 432
Ko J-H, Beers E, Han K-H (2006) Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*. Mol Genet Genom 276: 517-531
Koch MA, Weisshaar B, Kroymann J, Haubold B, Mitchell-Olds T (2001) Comparative genomics and regulatory evolution: conservation and function of the *Chs* and *Apetala3* promoters. Mol Biol Evol 18: 1882-1891
Kulheim C, Hui Yeoh S, Maintz J, Foley W, Moran G (2009) Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. BMC Genomics 10: 452
Kumar M, Thammannagowda S, Bulone V, Chiang V, Han K-H, Joshi CP, Mansfield SD, Mellerowicz E, Sundberg B, Teeri T, Ellis BE (2009) An update on the nomenclature for the cellulose synthase genes in *Populus*. Trends Plant Sci 14: 248-254
Legay S, Sivadon P, Blervacq A-S, Pavy N, Baghdady A, Tremblay L, Levasseur C, Ladouce N, Lapierre C, Séguin A, Hawkins S, Mackay J, Grima-Pettenati J (2010) *EgMYB1*, an R2R3 MYB transcription factor from *Eucalyptus* negatively regulates secondary cell wall formation in *Arabidopsis* and poplar. New Phytol 188: 774-786

Ling LL, Keohavong P, Dias C, Thilly WG (1991) Optimization of the polymerase chain reaction with regard to fidelity: modified T7, Taq, and vent DNA polymerases. Genome Res 1: 63-69
Livny J, Waldor MK (2009) Mining regulatory 5'UTRs from cDNA deep sequencing datasets. Nucleic Acids Res 38: 1504- 1514
Love J, Björklunda S, Vahala J, Hertzberg M, Kangasjärvib J, Sundberg B (2009) Ethylene is an endogenous stimulator of cell division in the cambial meristem of *Populus*. Proc Natl Acad Sci USA 106: 5984-5989
Lu S, Li L, Yi X, Joshi CP, Chiang VL (2008) Differential expression of three *Eucalyptus* secondary cell wall-related cellulose synthase genes in response to tension stress. J Exp Bot 59: 681-695
Maniatis T, Goodbourn S, Fischer JA (1987) Regulation of inducible and tissue-specific gene expression. Science 236: 1237-1245
McCarthy RL, Zhong R, Fowler S, Lyskowski D, Piyasena H, Carleton K, Spicer C, Ye Z-H (2010) The poplar MYB transcription factors, *PtrMYB3* and *PtrMYB20*, are involved in the regulation of secondary wall biosynthesis. Plant Cell Physiol 51: 1084-1090
Mellerowicz EJ, Sundberg B (2008) Wood cell walls: biosynthesis, developmental dynamics and their implications for wood properties. Curr Opin Plant Biol 11: 293-300
Mizrachi E, Hefer C, Ranik M, Joubert F, Myburg A (2010) De novo assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. BMC Genomics 11: 681
Molina C, Grotewold E (2005) Genome wide analysis of *Arabidopsis* core promoters. BMC Genomics 6: 25
Mutwil M, Debolt S, Persson S (2008) Cellulose synthesis: a complex complex. Curr Opin Plant Biol 11: 252-257
Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. Curr Opin Plant Biol 11(2):149-155.
Nei M (2007) The new mutation theory of phenotypic evolution. Proc Natl Acad Sci USA 104: 12235-12242
Nei M, Li W (1979) Mathematical model for studying genetic variation in term of restriction endonucleases. Proc Natl Acad Sci (USA) 76: 5267-5273
Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, Sederoff R, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9: 312
Pauli S, Rothnie HM, Chen G, He X, Hohn T (2004) The cauliflower mosaic virus 35S promoter extends into the transcribed region. J Virol 78: 12120-12128
Payn K, Dvorak W, Janse B, Myburg A (2008) Microsatellite diversity and genetic structure of the commercially important tropical tree species *Eucalyptus urophylla*, endemic to seven islands in eastern Indonesia. Tree Genet Genomes 4: 519-530
Piao SL, Fang JY, Ciais P, Peylin P, Huang Y, Sitch S, Wang T (2009) The carbon balance of terrestrial ecosystems in China. Nature 458: 1009-U1082
Popper ZA, Michel G, Herve C, Domozych DS, Willats WGT, Tuohy MG, Kloareg B, Stengel DB (2011) Evolution and diversity of plant cell walls: from algae to flowering plants. Annu Rev Plant Biol 62: 567-590
Pryor LD, Johnson LAS (1971) A classification of the Eucalypts. Australian National University, Canberra
Pyo H, Demura T, Fukuda H (2007) TERE; a novel *cis*-element responsible for a coordinated expression of genes related to programmed cell death and secondary wall formation during differentiation of tracheary elements. Plant J 51: 955-965

Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, Frederick WJ, Jr., Hallett JP, Leak DJ, Liotta CL, Mielenz JR, Murphy R, Templer R, Tschaplinski T (2006) The path forward for biofuels and biomaterials. Science 311: 484-489
Ranik M, Myburg AA (2006) Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. Tree Physiol 26: 545-556
Rathmann R, Szklo A, Schaeffer R (2010) Land use competition for production of food and liquid biofuels: An analysis of the arguments in the current debate. Renew Energ 35: 14-22
Regalbuto JR (2009) Cellulosic biofuels: Got gasoline? Science 325: 822- 824
Reineke AR, Bornberg-Bauer E, Gu J (2011) Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. Nucleic Acids Res 39: 6029-6043
Roberts AW, Bushoven JT (2007) The cellulose synthase (CESA) gene superfamily of the moss *Physcomitrella patens*. Plant Mol Bio 63: 207 - 219
Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497
Samuga A, Joshi CP (2004) Differential expression patterns of two new primary cell wall-related cellulose synthase cDNAs, *PtrCesA6* and *PtrCesA7* from aspen trees. Gene 334: 73-82
Sarkar P, Bosneaga E, Auer M (2009) Plant cell walls throughout evolution: towards a molecular understanding of their design principles. J. Exp. Bot. 60: 3615-3635
Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucl. Acids Res. 18: 6097-6100
Shi R, Sun Y-H, Li Q, Heber S, Sederoff R, Chiang VL (2010) Towards a systems approach for lignin biosynthesis in *Populus trichocarpa*: transcript abundance and specificity of the monolignol biosynthetic genes. Plant Cell Physiol 51: 144-163
Steane DA, McKinnon GE, Vaillancourt RE, Potts BM (1999) ITS sequence data resolve higher level relationships among the eucalypts. Mol Phylogenet Evol 12: 215-223
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739
Tanaka K, Murata K, Yamazaki M, Onosato K, Miyao A, Hirochika H (2003) Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. Plant Physiol 133: 73-83
Tanaka T, Koyanagi KO, Itoh T (2009) Highly diversified molecular evolution of downstream transcription start sites in rice and *Arabidopsis*. Plant Physiol 149: 1316 - 1324
Tanay A, Regev A, Shamir R (2005) Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. Proc Natl Acad Sci (USA) 102: 7203-7208
Taylor NG (2008) Cellulose biosynthesis and deposition in higher plants. New Phytol 178: 239-252
Taylor NG, Gardiner JC, Whiteman R, Turner SR (2004) Cellulose synthesis in the *Arabidopsis* secondary cell wall. Cellulose 11: 329-338
Taylor NG, Howells RM, Huttly AK, Vickers K, Turner SR (2003) Interactions among three distinct CESA proteins essential for cellulose synthesis. Proc Natl Acad Sci (USA) 100: 1450-1455
Than C, Ruths D, Nakhleh L (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics 9: 322
Thomas-Chollier M, Sand O, Turatsinze J-V, Janky Rs, Defrance M, Vervisch E, Brohee S, van Helden J (2008) RSAT: regulatory sequence analysis tools. Nucleic Acids Res 36: W119- W127

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680
Tjaden G, Edwards JW, Coruzzi GM (1995) Cis elements and trans-acting factors affecting regulation of a nonphotosynthetic light-regulated gene for chloroplast glutamine synthetase. Plant Physiol 108: 1109-1117
Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137-144
Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y (2009) Unraveling transcriptional control in *Arabidopsis* using cis-regulatory elements and coexpression networks. Plant Physiol 150: 535- 546
Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256-276
Wijaya E, Yiu S-M, Son NT, Kanagasabai R, Sung W-K (2008) MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders. Bioinformatics 24: 2288-2295
Wu A-M, Hu J, Liu J-Y (2009) Functional analysis of a cotton cellulose synthase A4 gene promoter in transgenic tobacco plants. Plant Cell Rep 28: 1539-1548
Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J (2007) Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. Nucleic Acids Res 35: 6219-6226
Yazaki J, Kishimoto N, Nagata Y, Ishikawa M, Fujii F, Hashimoto A, Shimbo K, Shimatani Z, Kojima K, Suzuki K, Yamamoto M, Honda S, Endo A, Yoshida Y, Sato Y, Takeuchi K, Toyoshima K, Miyamoto C, Wu J, Sasaki T, Sakata K, Yamamoto K, Iba K, Oda T, Otomo Y, Murakami K, Matsubara K, Kawai J, Carninci P, Hayashizaki Y, Kikuchi S (2003) Genomics approach to abscisic acid- and gibberellin-responsive genes in rice. DNA Res 10: 249-261
Yin Y, Huang J, Xu Y (2009) The cellulose synthase superfamily in fully sequenced plants and algae. BMC Plant Biol 9: 99
Zhao L, Lu L, Zhang L, Wang A, Wang N, Liang Z, Lu X, Tang K (2009) Molecular evolution of the E8 promoter in tomato and some of its relative wild species. J Biosci 34: 71-83
Zhong R, Lee C, Ye Z-H (2010) Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. Trends Plant Sci 15: 625-632
Zhong R, Lee C, Zhou J, McCarthy RL, Ye ZH (2008) A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. Plant Cell 20: 2763- 2782

Data Archiving Statement

All sequences used for analyses have been deposited on the GenBank database at NCBI (http://www.ncbi.nlm.nih.gov/) or have been previously published and are referenced as such.

The Genbank accession numbers are listed in full in Online Resource 5 and in the Materials and Methods. The accession numbers are: JN573683 - JN573751 and JN573752 - JN573783.