Identification of a *Cis*-acting Regulatory Polymorphism in a Eucalypt *Cobra*-like Gene Affecting Cellulose Content


*CSIRO Plant Industry, GPO Box 1600, Canberra, ACT 2601 Australia

**CSIRO Sustainable Ecosystems, Private Bag 12, Hobart, Tasmania 7001 Australia

§Present address: Key Laboratory for Genetics and Breeding in Forest Trees and Ornamental Plants, Beijing Forestry University, Beijing 100083, P. R. China

Sequence data from this article has been deposited with the GenBank Data Library under the accession no FJ213604. (EniCOBLAA)
Running Title: Regulatory polymorphism in EniCOBL4A

Corresponding author:

Bala R Thumma

CSIRO Plant Industry, GPO Box 1600, Canberra, ACT 2601 Australia

Ph: +61 2 62464856; Fax: +61 2 62464864

Reddy.thumma@csiro.au

Key words: Single nucleotide polymorphism, Allelic expression imbalance, Methylation, Linkage disequilibrium, Association studies.
ABSTRACT

Populations with low linkage disequilibrium (LD) offer unique opportunities to study molecular variants influencing quantitative traits. We exploited the low LD in forest trees to identify functional polymorphisms in a *Eucalyptus nitens* COBRA-like gene (*EniCOBL4A*), whose *Arabidopsis* homolog has been implicated in cellulose deposition.

Linkage analysis in a full-sib family revealed that *EniCOBL4A* is the most strongly associated marker in a quantitative trait locus (QTL) region for cellulose content. Analysis of LD by genotyping 11 common single nucleotide polymorphisms (SNPs) and a simple sequence repeat (SSR) in an association population revealed that LD declines within the length of the gene. Using association studies we fine-mapped the effect of the gene to SNP7, a synonymous SNP in exon 5, which occurs between two small haplotype blocks.

We observed patterns of allelic expression imbalance (AEI) and differential binding of nuclear proteins to the SNP7 region that indicate that SNP7 is a cis-acting regulatory polymorphism affecting allelic expression. We also observed AEI in SNP7 heterozygotes in a full-sib family which is linked to heritable allele-specific methylation near SNP7. This study demonstrates the potential to reveal functional polymorphisms underlying quantitative traits in low LD populations.
INTRODUCTION

In plants, gene function is typically investigated by reverse genetic approaches such as knockout mutants and overexpression transgenics. Functional analysis of subtle allele effects under native conditions is difficult. The large extent of linkage disequilibrium (LD) is one of the obstacles to fine mapping of functional variants in many plant species. In this context forest trees offer unique opportunities for fine mapping of candidate genes and functional analysis of gene variants, as LD and population structure are generally low (BROWN et al. 2004; NEALE 2007; THUMMA et al. 2005). In inbreeding species LD typically extends over much larger distances compared to outcrossing species (NORDBORG 2000). In Arabidopsis, average LD extends to about 10kb (KIM et al. 2007), while in the outcrossing maize and forest tree species, rapid decline in LD (within 1kb and lower) was observed in a number of candidate genes (SAVOLAINEN and PYHAJARVI 2007; THORNSBERRY et al. 2001). Some of the factors that are likely to have contributed to the low LD in forest trees include large effective population size, their outcrossing habit and long history of recombination.

The functional significance of non-synonymous polymorphisms may be linked to protein phenotype while synonymous and non-coding SNPs can affect phenotype by impacting alternative splicing, altering mRNA stability and/or allelic expression imbalance (AEI). Allelic expression imbalance or allele-specific expression analysis in heterozygous individuals is a powerful method for identifying cis-acting regulatory variants. Expression of a particular allele is controlled by both cis-acting factors such as DNA polymorphisms and methylation and trans-acting factors such as transcription factors. Differential expression of two alleles within an individual, where they are exposed to the same cellular environment, indicates the presence of cis-acting regulatory polymorphisms (TAO et al. 2006). While total gene expression is influenced by both cis- and trans-acting factors and
environmental factors, allele-specific expression is primarily affected by cis-acting factors. Analysis of AEI is therefore more robust compared to total gene expression. Since regulatory polymorphisms present in promoters or within genes (i.e. enhancers) will show an effect on allele expression levels, these can be identified by measuring AEI (Pastinen and Hudson 2004).

Many of the AEI studies in humans have been successful in identifying cis-regulatory haplotypes underlying gene expression (Pastinen et al. 2005; Tao et al. 2006). However, identification of the functional polymorphisms controlling allelic expression is generally not possible because of the extent of LD in humans. Further analysis to identify functional variants is generally carried out using transient expression studies. Several AEI studies have been reported in maize (Guo et al. 2004; Springer and Stupar 2007; Stupar and Springer 2006). The proportion of cis-regulatory variation affecting AEI was higher than trans-regulatory variation and evidence of tissue-specific differences in AEI was detected (Springer and Stupar 2007). However, all these studies in maize were primarily aimed at understanding the genetic control or mode of inheritance of regulatory variation. In maize AEI has been used to identify upstream regions containing regulatory factors (Clark et al. 2006; Salvi et al. 2007); however, there are no studies in plants where AEI was used to identify regulatory polymorphisms affecting quantitative trait variation.

Identification of genes and gene variants controlling wood quality traits is an important objective in many forest tree breeding programs, as small fraction of a percentage point improvements in traits such as pulp yield, can deliver large gains for a pulp mill. During studies of gene expression and wood properties in Eucalyptus nitens branches (Qiu et al. 2008) we identified two COBRA-like ESTs that share strongest DNA sequence similarity with AtCOBL4. These genes were differentially expressed in upper (tension wood) and lower (opposite wood) sides of branch xylem where large differences in
cellulose content and cellulose microfibril orientation were observed. In *Arabidopsis*

*Cobra* and *Cobra*-like genes have been implicated in cellulose synthesis and deposition
in secondary cell walls (SCHINDELMAN *et al.* 2001). Reverse genetic studies in *Arabidopsis*
have shown that insertions in *(COBL4)* resulted in a large decrease in cellulose content
(BROWN *et al.* 2005). Mutations in genes homologous with *(COBL4)* in rice (*BC1*) (LI *et al.*
2003) and maize (*bk2*) (CHING *et al.* 2006) caused reductions in cellulose content in
secondary cell walls of stems.

In this study we exploited the low LD found in *(COBL4)* gene from *E. nitens* to fine
map functional polymorphisms. We present a range of experimental evidence to show that
SNP7 is a cis-acting regulatory polymorphism that affects cellulose content in *E. nitens.*
MATERIALS AND METHODS

Populations, phenotypes and genotypes: The association population comprised 420 nine-year-old *E. nitens* growing at Meunna, north western Tasmania (41° 06’S 145° 28’E). Each tree was derived from an unrelated open pollinated family (consisting of half sibs) growing in the central highlands of Victoria, from where most of the Australian *E. nitens* breeding material is sourced. The three generation full-sib family comprised 290 eight-year-old trees grown at West Ridgley, north western Tasmania (41° 14’S 145° 78’E). Two of the grand parents came from the central Victorian highlands and two came from New South Wales. Wood cores were collected from all trees at 1.3m from ground level in 2002. Cellulose microfibril orientation and density were measured using Silviscan II (EVANS and ILIC 2001). Cellulose and lignin content and kraft pulp yield were predicted from ground powder from whole cores or across intact cores by near-infrared (NIR) spectra using appropriate calibration models in the methods of Schimleck et al. (2004). Details of the method describing cellulose estimation based on NIR spectroscopy are presented in Appendix S1. Leaves collected from each tree were used for DNA isolation. The entire *EniCOBL4A* gene, including 1.0 kb of promoter region was sequenced in 18-20 individuals to identify common SNPs (MAF>0.10). Common SNPs were genotyped in 300 trees in the association population using the multiplex ligation-dependent probe amplification (MLPA) assay (SCHOUTEN et al. 2002). Accuracy of the genotype calls for SNP7 was tested using primer extension in 300 trees. One SNP from the promoter, SNP1 from the 5’UTR, SNP2 from exon 1 and SNP7 from exon 5 were genotyped using the GoldenGate (Illumina) (FAN et al. 2003) bead assay. SNP7 was genotyped in the entire population of 420 trees with the primer extension method and GoldenGate assay. SNP7 and SNP13 from the 3’UTR were
genotyped in the grand parents, parents and across 290 individuals from full-sib family using the primer extension method.

**Phylogenetic analysis:** The *EgrCOBL4A* and *EgrCOBL4B* genes were identified in a *E. grandis* xylem cDNA library (Qiu et al. 2008). We used the EST sequence from *Egr-COBL4A* to isolate full-length gene in *E. nitens* and 5'UTR and promoter regions of *EniCOBL4A* were obtained using Genome Walker (Clontech). Protein sequences were aligned using ClustalX (Thompson et al. 1994) and the alignments used as input into MEGA 3.1 (Kumar et al. 2004) to generate an unrooted phylogenetic tree. All analyses were performed using the default parameters. Statistical confidence of the nodes of the tree is based on 1000 bootstrap replicates.

**Allelic expression imbalance:** Allelic expression imbalance (AEI) was measured in heterozygous trees using the primer extension method (Gábor Mátyás et al. 2002; Yan et al. 2002). RT-PCR was performed on DNAse treated total RNA extracted from xylem tissue (Southern et al. 1998) using the forward primer 5'-GTCGGCTGTCTGGTACTTC-3' and reverse primer 5'-CCCGTGTCATTTATGGATTC-3' to amplify a portion of the gene containing SNP7. The reverse primer spanned exon 5 and exon 6 to minimise amplification from genomic DNA. Allele-specific primer extension products were separated by capillary electrophoresis and peak height was used to measure the relative ratio of each allele. Allelic ratios of the cDNA samples were standardised with mean allelic ratios obtained using genomic DNA (Zhang et al. 2005). Four technical replicates were performed for each sample. In the association population AEI estimation with SNP7 was done using two biological replicates. Xylem samples for RNA were collected in the mornings over several days at each trial site.
**Electrophoretic mobility shift assay (EMSA):** Nuclear proteins were extracted from xylem tissue of four unrelated trees using a plant nuclear extraction kit (Sigma). Two biotin-labelled 25 mer oligonucleotides spanning SNP7 were synthesised and annealed to their complementary oligonucleotides using a thermocycler. Probe-protein binding reactions were performed at room temperature using a LightShift chemiluminescent EMSA kit (Pierce Biotechnology). We incubated 2 ul (about 3 ug) of crude protein and 1 pmol biotin labelled probe together with 2 ul of binding buffer, 2 ul of poly di/dc, 1 ul of glycerol and 1 ul of MgCl₂ at room temperature for 20 minutes. The binding reaction was loaded onto a 6% polyacrylamide gel in 0.5X TBE and electrophoresed for 20-30 minutes. Probe-protein complexes were transferred to a nylon membrane and developed using a LightShift biotin detection kit (Pierce Biotechnology).

**Methylation analysis:** The methylation status of CpG dinucleotides in AA and GG homozygous trees from the full-sib family were examined by preparing bisulfite treated DNA using a Methyl Code kit (Invitrogen). Primers were designed using MethPrimer software (Li and Dahya 2002) to amplify 250 bp surrounding SNP7. PCR amplification products were cloned and about 15 clones sequenced to reveal methylation status. Methylation-sensitive single nucleotide primer extension (Ms-SNuPE, Gonzalgo and Jones 2002) was used to genotype sites which were found to be methylated and to interrogate other CpG sites for their methylation status.

**Statistical analyses:** We used ‘SAS genetics’ to test Hardy-Weinberg Equilibrium (HWE) of the SNPs and to estimate LD using genotypic data. Haplotypes from genotype data were estimated and assigned to each tree using the ‘PHASE’ program (Stephens et al. 2001). Statistical analysis of genotype classes showing AEI was done using ‘Graphpad’
Prism’ software. We used single marker as well as interval mapping for QTL analysis in the full-sib family. We used the GLM function of ‘TASSEL’ (BRADBURY et al. 2007) software and SAS to identify SNP-trait associations in association population. Seven traits were tested for association with 11 SNPs. Significance parameters were based on 1000 permutation tests. Experiment-wise $p$-values were calculated based on the minimum $p$-value across all tests from permuted trait data compared to the original $p$-value for each marker. Adjusted $P$-values representing the percentage of times the permuted $P$-value was lower than the original $P$-value were used in testing for significant associations. Haplotype association tests were done on three marker sliding window using haplotype trend regression software ((ZAYKIN et al. 2002). Overall significance of the haplotype was based on 1000 permutation tests. SNP7 associating with traits were further analysed in additive, dominant and recessive models using the ‘SNPStats’ program (SOLE et al. 2006).
RESULTS

Eucalyptus has at least two genes homologous to AtCOBL4: *E. nitens* homologs (*EniCOBL4A* and *EniCOBL4B*) of ESTs identified in earlier studies (Qiu et al. 2008), share strongest amino acid identity (81% and 84%, respectively) with AtCOBL4. Phylogenetic analysis revealed that the two eucalypt genes reside in a subclade that includes AtCOBL4 and its homologs from other plants (Figure 1).

Nucleotide diversity and linkage disequilibrium in *EniCOBL4A*: A 3.3 kb region containing the *EniCOBL4A* gene and 830 bp upstream of the predicted transcription start site was sequenced in 18 unrelated individuals. Eleven common SNPs (freq. > 0.1) were detected, 5 from coding and 6 from non-coding regions (Figure 2). Of the coding-region SNPs, SNP2 in exon 1 and SNP4 in exon 3 were non-synonymous. Within the predicted promoter region, we identified a SNP (PrSNP) at -665 bp and a tri-nucleotide microsatellite (PrSSR) at -365 bp.

Tests of neutrality with allele sequences from 18 trees revealed little evidence of selection at the *EniCOBL4A* locus. Analysis of the minimum number of recombination events using the four-gamete test (Hudson and Kaplan 1985) identified two potential regions of recombination; one between SNP1 and SNP3 and one between SNP7 and SNP8. LD analysis using genotype data from 300 trees in the association population (Figure 2) showed that PrSSR and SNP2 had no significant LD with any SNPs and revealed two distinct haplotype blocks from SNP3 to 6 (545 bp) and SNP8 to 10 (395 bp). Within each block, LD among the SNPs was high ($r^2 > 0.80$) while between the two haplotype blocks LD was low ($r^2 < 0.30$). SNP7 from exon 5, which occurred between the two haplotype blocks, had low LD with SNPs from the two blocks. Haplotype analysis with genotype data
from all SNPs revealed thirteen haplotypes that occurred in 95% of trees in the association population (Table 1).

Haplotype structure in the full-sib family was determined by genotyping all the SNPs in the grandparents, parents and progeny. This revealed that the progeny inherited three haplotypes from the parents (Figure 3). Haplotypes 1, 2 and 3 in the full-sib family correspond to haplotypes 12, 1 and 13 respectively, in the association population (Table 1). Haplotypes 1 and 3 in the full-sib family are very rare in the association population (Freq < 0.01).

**EniCOBL4A co-locates to a QTL region for cellulose content:** We tested for association of *EniCOBL4A* with chemical (lignin content, cellulose content, pulp yield) and physical (cellulose microfibril angle and wood density) wood quality traits in the *E. nitens* full-sib family. Since SNP7 was heterozygous in both the parents of the full-sib family we used this SNP to genotype the 275 progeny. Analysis of variance (ANOVA) revealed a significant association of *EniCOBL4A* with cellulose content and kraft pulp yield (Table 2). AA homozygotes were associated with higher trait values while GG homozygotes were associated with lower trait values. Further analysis of the data assuming additive, dominant and recessive models revealed the strongest association between *EniCOBL4A* and cellulose and kraft pulp yield was observed under a dominant model (Table 2). Additional tests in the full-sib family with four genotype classes as determined by SNP10 (Figure 3) increased the significance of the *EniCOBL4A*-cellulose association compared to the codominant model (Figure 4a). Linkage analysis with several genetic markers revealed that *EniCOBL4A* co-locates to a QTL region for cellulose content on linkage group 10 of an *E. nitens* linkage map (BYRNE et al. 1995). Within the QTL region the significance of the *EniCOBL4A* association with cellulose content was higher than the other linked markers (Figure 4b).
SNP7 of *EniCOBL4A* is associated with cellulose content and kraft pulp yield: To fine map *EniCOBL4A* we initially genotyped all 11 SNPs in 300 trees from the *E. nitens* association population. Earlier studies have shown that there is no significant structure in the population used in the present study (Thumma et al. 2005). Single marker analysis with seven wood traits revealed that SNP7 from exon5 was significantly associated with cellulose content and kraft pulp yield (Table 3). Haplotype analysis did not improve the significance of the SNP7 association suggesting that the effect is restricted to the SNP7 region (Table 3). A further 116 samples from the association population were genotyped for SNP7 alone and associations with cellulose and kraft pulp yield for the entire population are presented in Table 4. The frequency of AA homozygotes and AG heterozygotes for SNP7 was 0.45 and the GG homozygote was 0.1 in the association population. Mean cellulose values of the AA homozygotes were highest (40.85 %) and those of GG homozygotes were lowest (40.44 %), while mean cellulose of AG heterozygotes were intermediate (40.58 %). Similarly, mean kraft pulp yield values of AA homozygotes were higher (52.60 %) than those of GG homozygotes (52.31 %) and AG heterozygotes (52.29 %). Further analysis of the effect of SNP7 under different genetic models identified strong evidence of association between SNP7 and cellulose under additive and/or dominant genetic models (Table 4). Similarly, a strong association was observed with kraft pulp yield under a dominant model (Table 4).

While the absolute values of cellulose content and kraft pulp yield differ between the association population and the full-sib population, the AA homozygotes were associated with higher trait values while GG homozygotes were associated with lower trait values in both the populations. Also, the trait differences between AA and GG homozygotes for cellulose content and kraft pulp yield were similar in the two populations,
varying by 0.4 and 0.3 % respectively (Tables 2 & 4). However, while the mean trait values of the heterozygotes were lower compared to AA homozygotes in the association population, they were higher in the full-sib family. Similarly, allele-G was dominant in the association population, in contrast to the full-sib family where allele-A was generally dominant (Tables 2 & 4).

**SNP7 alleles of EniCOBL4A are differentially expressed:** To test its functional significance we used SNP7 (exonic) to examine allele-specific expression using xylem cDNA from 16 unrelated trees heterozygous at SNP7. Allelic expression ratio estimates had very low standard errors suggesting that these estimates are robust and repeatable (Figure 5). In all trees expression of allele-A was lower than allele-G. Allele-A is expressed at about 85 % of the level of allele-G with the A/G expression ranging from 0.67 to 0.98 (Figure 5a). Three distinct genotype classes are predicted with high probability (100 %) for 15 of the 16 trees when genotype information for SNP2 to SNP11 is used in phase analysis. Population frequencies of genotype class 1 and genotype class 2 are 0.15 while that of genotype class 3 is 0.08. AEI appeared to be related to the haplotype structure of the trees (Figure 5a). AEI of genotype class 1 (GTGTCG/GCACGAGC) ranged from 0.67 to 0.82 (mean 0.74), genotype class 2 (GTGTCG/GCACGAGC) ranged from 0.78 to 0.98 (mean 0.89) and genotype class 3 (ATGTCA/GTC/GTC) had an AEI of about 0.97. Allelic expression of genomic DNA (gDNA) controls ranged from 0.9 to 1.1 with a mean of 1.0. Consistent lower expression of allele-A in all 16 SNP7 heterozygous trees suggests that SNP7 influences AEI.

**SNP7 contributes to the regulation of allelic expression imbalance:**

We directly tested the impact of SNP7 on allelic expression by comparing AEI in trees of genotype class 1 with a tree of the same genotype except that it was homozygous at SNP7 (GTGTCAGTC/GCACGAC; Figure 5b); designated as genotype class 1A. We used
SNP4 for measuring AEI in this tree after confirming SNP4 AEI measurements were similar to SNP7 ($r^2 = 0.95$). We measured AEI in an additional five genotype class 1 trees using SNP4 and found it ranged from 0.62 to 0.77 with a mean of 0.69 while AEI of the tree with genotype class 1A was 0.91 (Figure 5b).

We tested if AEI in genotype class 1 is significantly different from gDNA controls and other genotype classes. Even though we had a single tree belonging to genotype class 1A, we were able to use the information from the two alleles in a statistical analysis similar to Zhang et al. (2005). Non-parametric (Kruskal-Wallis) analysis with Dunn’s post test showed that allelic expression in genotype class 1 was significantly different from the gDNA controls ($P < 0.0001$) implying unequal expression of the two alleles in this genotype class, whereas other genotype classes, including class 1A were not significantly different from gDNA. Equal expression of the two alleles in a tree belonging to genotype class 1A provides further support that SNP7 plays a role in the regulation of AEI. However equal expression of two alleles in genotype classes 2 and 3 indicate other variants interact with SNP7.

**Nuclear proteins bind differentially to SNP7 alleles:** To test whether SNP7 is part of a *cis*-acting regulatory element within the *EniCOBL4A* gene we carried out electrophoretic mobility shift assays (EMSA) using two 25 bp biotin-labelled probes spanning the SNP7 region but differing by the presence of an A or G nucleotide at SNP7 (Figure 6). While nuclear proteins were found to bind to both alleles, we observed greater binding of nuclear protein to allele-A compared to allele-G (Figure 6a). Two bands were observed in the lanes with mobility shifts suggesting two different protein complexes were binding to the probes. These experiments were repeated several times and with different concentrations of probes and we observed consistent differences in the intensity of the shifted bands between the two probes (Figure 6b).
**Allelic expression imbalance in the full-sib family:** We examined AEI in 28 heterozygous trees in a full-sib family whose parents were both heterozygous at SNP7. Using SNP10 (Figure 3) we separated the two heterozygote classes as 1,4 and 2,3. Class 1,4 heterozygotes had a mean expression ratio (A/G) of 0.97 while 2,3 heterozygotes had a mean expression ratio of 1.29 (Figure 7). Allelic expression in class 1,4 heterozygotes was not significantly different from gDNA controls (non-parametric analysis with Dunn’s post test $P > 0.05$) while in class 2,3 heterozygotes allelic expression was significantly different from gDNA controls ($P < 0.0001$) indicating differential expression of the two alleles in 2,3 heterozygotes. In contrast to AEI in the association population, expression of allele-A was higher than allele-G in the 2,3 heterozygote class of the full-sib family. Class 2,3 heterozygotes which showed AEI, have significantly higher cellulose content compared to 1,4 heterozygotes (Figure 4a, $P = 0.018$), suggesting that cellulose content and the level of allelic expression of allele-A are positively correlated. The association of allelic expression ratios with specific genotype classes in the association population (Figure 5) and full-sib family (Figure 7) indicate that AEI in both populations is heritable.

**Allelic expression imbalance in the full-sib family is linked to allele-specific methylation:** Since allele-G of SNP7 occurs in a CpG dinucleotide, we sequenced bisulfite treated genomic DNA from an AA homozygous and a GG homozygous tree from the full-sib family to reveal methylated sites in a 250 bp region surrounding SNP7. We did not detect methylation of allele-G at SNP7; however, we detected methylation specific to the AA homozygous tree at a single CpG dinucleotide in intron 4 about 75 bp upstream of SNP7. Approximately half of the clones sequenced from the AA homozygous tree had a C-nucleotide at the methylated site and the other half had a T-nucleotide suggesting that the tree is heterozygous for methylation. To study the heritability of the methylated site in the full-sib family, we analysed the grandparents, parents and four
progeny trees belonging to four different genotype classes using methylation sensitive
single nucleotide primer extension (MS-SNuPE) (GONZALGO and JONES 2002). This
analysis revealed that the methylation of the CpG site is heritable and allele-specific in the
full-sib family (see inheritance of methylated allele in Figure 3). One of the grandparents of
the paternal parent (SN) is heterozygous for methylation. The paternal parent inherited the
methylated allele and passed it on to the progeny in an allele-specific manner. Allele-A of
the 2,3 heterozygous class (Figure 3) in the progeny, which showed higher expression, was
methylated; while in the 1,4 heterozygous class allele-A was unmethylated. Methylation,
therefore, is associated with the allele which showed higher expression. We observed the
same allele-specific methylation pattern in DNA from either xylem or leaf tissue,
suggesting that methylation of this site is not tissue-specific. Methylation of the intron 4
site is rare in the association population as analysis of the DNA of over two hundred trees
using primer extension revealed only two trees that were methylated. In both populations,
methylation at other CpG dinucleotides between SNP6 and SNP8 was absent.
In this study we exploited the low LD usually found in forest trees to fine map functional polymorphisms in *EniCOBL4A*. To identify *cis*-acting regulatory polymorphisms we analysed AEI in an association population and a full-sib family and carried out EMSA analysis. Regulatory mechanisms are most often functionally characterised by *in vitro* methods such as transient transfection experiments. The main drawback of such approaches is that the experiments are performed away from the normal chromosomal environment and may not represent the *in vivo* condition (Pastinen *et al.* 2004). Analysis of AEI *in vivo* using SNP markers from transcribed regions is a powerful approach to detect regulatory polymorphisms (Knights 2005). In a recent report Cirulli and Goldstein (2007) tested the congruence between reporter gene assays and *in vivo* methods of allelic expression. They showed that *in vitro* assays fail to predict *in vivo* effects of regulatory polymorphisms suggesting that the affects of polymorphisms on gene expression are more complex and context dependent.

Differential expression of alleles in a heterozygous individual where both alleles are exposed to the same environmental and *trans*-acting factors strongly suggests the presence of *cis*-regulating polymorphisms. Therefore AEI can reveal subtle *cis*-acting differences even when feedback mechanisms dilute the total expression differences (Pastinen *et al.* 2006). Using SNP7 in the association population we found that allele-A is expressed consistently at a lower level than allele-G in all the trees tested even though the expression of the allele A and G are not significantly different in genotype 2 and 3. If expression imbalance consistently shifts in one direction i.e. one allele is expressed more strongly than the other allele in all the samples, this suggests that the transcribed marker used for studying AEI itself could be the *cis*-regulating polymorphism or in high LD with such a
polymorphism (TAO et al. 2006). As SNP7 occurs in low LD with other markers from this
gene this suggests that SNP7 itself may be the cis-regulating polymorphism controlling
AEI. Allelic expression analysis in an individual tree containing the two differentially
expressed haplotypes but homozygous at SNP7 revealed similar expression between the
two haplotypes. Taken together these results strongly suggest that SNP7 is a regulatory
polymorphism giving rise to the AEI. This is further supported by EMSA analysis which
revealed preferential binding of nuclear proteins to allele-A at SNP7. In many human
 genetic studies regulatory haplotypes controlling AEI were identified by LD mapping
 (KNIGHT et al. 2003; PASTINEN et al. 2005; TAO et al. 2006). However, further fine
mapping of these haplotypes to identify cis-regulatory polymorphisms was not possible
because of the extent of LD in humans. We were able to fine map the control of AEI to a
region surrounding SNP7 because of the low LD observed in the gene. Searching through
the plant cis-acting regulatory DNA elements (PLACE) database (KENICHI HIGO et al.
1999), we found two ‘CAAT’ box elements near SNP7. One was two bp upstream of SNP7
and the other was nine bp downstream of the SNP7. CAAT box elements are responsible
for tissue-specific promoter activity of pea legumin gene in tobacco (SHIRSAT et al. 1989).

Our discovery of a synonymous exonic SNP (SNP7) lying in a cis-acting regulatory
element is somewhat surprising. Regulatory elements such as transcriptional enhancers and
silencers are generally found in non-coding sequences such as promoters, introns and
untranslated regions. However there are a few reports of the occurrence of transcriptional
regulatory elements in coding regions (GEORGINA LANG 2005; NEZNAHOV et al. 1997;
PIERCE et al. 2006). Barthel and Liu (2008) used computational approaches to identify a
transcriptional regulatory element in exon 2 of ADAMTS5, a human myogenesis gene.
Using different approaches such as chromatin immunoprecipitation and reporter gene
assays they confirmed that the regulatory element in exon 2 acts as a transcriptional
enhancer. Tao et al. (2006) used allele-specific expression, EMSA and reporter gene assays to identify several polymorphisms in the human KRT gene which are involved in transcriptional regulation including a coding region polymorphism. High resolution mapping of expression QTLs (eQTLs) in humans by Veyrieras et al. (2008) revealed that SNPs in exons are two fold more likely to be eQTNs than SNPs in introns.

In the full-sib family we detected heritable, allele-specific methylation of allele-A 75 bp upstream of SNP7 which is associated with differential allelic expression of 2,3 heterozygotes. There are few reports in plants where allele-specific methylation (ASM) was associated with AEI of nonimprinted genes. In a survey of genome wide methylation patterns in humans, Kerkel et al. (2008) identified sequence dependant ASM at several nonimprinted genes. At two loci they found association of AEI with sequence-dependant ASM. They suggested that the presence of ASM is an indicator of nearby regulatory polymorphisms and identification of ASM could be used for fine mapping functional variants (Kerkel et al. 2008). However, the association between allelic expression and methylation observed in the full-sib family requires further functional analyses to understand the role of methylation in allelic expression. Methylation is unlikely to be involved in the AEI observed in the association population as methylation at the intron 4 CpG site is rare.

In all the trees measured for AEI in the association population expression of allele-A is either lower or equal to that of allele-G. In contrast, in the full-sib family expression of allele-A is either equal or higher than that of allele-G. In the association population, where expression of allele-A is low, methylation at intron 4 CpG site is very rare. Higher expression of allele-A in the full-sib family, therefore, may be linked to its methylation. Methylation of CpG islands in promoter regions is usually associated with lower expression; however, the relationship between gene expression and transcribed region
methylation is not clear. Genome-wide methylation analysis of *Arabidopsis* revealed that more than 30% of genes are methylated in transcribed regions and the methylated genes tended to be more highly expressed (Zhang et al. 2006). Further evidence for higher expression of methylated genes comes from genetic studies of human diseases (Polesskaya et al. 2006; Unoki and Nakamura 2003). Recent studies suggest that cyclic methylation and demethylation of CpG sites in promoter regions can play a role in transcriptional regulation of genes (Kangaspeska et al. 2008; Metivier et al. 2008). Such transient demethylation may underlie the higher expression of methylated genes observed in these studies.

Quantitative trait locus mapping revealed that EniCOBL4A maps to the region of a QTL for cellulose in LG10 of an *E. nitens* linkage map (Figure 4b). Association mapping with a number of SNPs within *EniCOBL4A* revealed that SNP7 is significantly associated with cellulose content and kraft pulp yield (Tables 3 & 4). This observation is consistent with data from *Arabidopsis* and other plant species where COBL4 has been shown to be involved in cellulose deposition and cell wall formation. This study demonstrates the potential for high resolution mapping of candidate genes in a species with low LD. We observed several rare SNPs and eleven common SNPs in *EniCOBL4A* in 18 unrelated trees. In contrast, we did not detect any SNPs in the AtCOBL4 gene in 18 *Arabidopsis* ecotypes in a published SNP database (http://natural.salk.edu). This difference likely reflects fundamental differences in the life histories of the two species with *Arabidopsis* being highly inbred compared to eucalypts which are mostly out-crossing. While the capacity for fine mapping of candidate genes as described here may be limited in *Arabidopsis*, genome-wide association studies should result in high resolution mapping of genomic regions containing a few candidate genes.
Allele-A was found to be dominant in the full-sib family while allele-G was dominant in the association population. This may be due to epistatic effects or it may be due to the differences in the genetic background of the two populations with the full-sib family carrying much less diversity than the association population. It may also be related to the contrasting patterns of allelic expression observed in the two populations. In the full-sib family mean allelic expression of allele-A is higher than allele-G (AEI = 1.12; Figure 7); however, in the association population mean allelic expression of allele-A is lower than allele-G (AEI = 0.84; Figure 5). Similarly there appears to be a correlation between allelic expression and performance of the heterozygotes in the two populations. In the full-sib family, AG heterozygotes are associated with high trait values whereas in the association population they are associated with low trait values (Tables 2 and 4). Moreover, within the full-sib family where the two heterozygote classes showed different allelic expression patterns, there was a significant difference in the cellulose content with 2,3 heterozygotes having higher trait values compared to 1,4 heterozygotes (Figure 4a). It is tempting to speculate that higher expression of allele-A in the full-sib family may give rise to the higher trait values of 2,3 heterozygotes. Equal expression of both alleles in 1,4 heterozygotes may lead to high trait values similar to AA homozygotes, while lower expression of allele-A may explain the lower trait values of AG heterozygotes in the association population. This hypothesis is supported by studies in maize which showed that genes with differential allelic expression tended to be correlated with yield heterosis but not genes with high expression (Guo et al. 2006).

Recently Serre et al (2008) used an Illumina BeadArray to analyse AEI in more than 1000 genes. Development of high throughput methods for assessing AEI will be helpful in selection of candidate genes for association studies. While cis-acting affects may be revealed by AEI, important trans-acting effects may not be identified by studying AEI
alone. Kirst et al (2005; 2004) using eQTL analysis in Eucalyptus found that genes involved in the lignin biosynthetic pathway are predominantly regulated through cis-acting effects; however, a number of genes in differentiating xylem tissue are also regulated through trans-acting effects. Combination of AEI studies with eQTL studies should therefore be helpful in identification of both cis and trans acting regulatory effects. In conclusion this study demonstrates the opportunities for identifying functional polymorphisms in populations with low LD. The observation that an epigenetic polymorphisms is likely to influence AEI and subsequently phenotypic variation in cellulose content, suggests epigenetic variation may be a relatively common mechanism underlying quantitative trait variation.
We thank Forestry Tasmania and Gunns Ltd for access to the *E. nitens* association and full-sib mapping populations respectively. We acknowledge financial support from Forest and Wood Products Australia (grant number: PNO7.3024).
LITERATURE CITED


STUPAR, R. M., and N. M. SPRINGER, 2006 Cis-transcriptional Variation in Maize Inbred Lines B73 and Mo17 Leads to Additive Expression Patterns in the F1 Hybrid. Genetics 173: 2199-2210.


Table 1. Haplotype frequencies in the association population estimated using genotypic information from 11 SNPs

<table>
<thead>
<tr>
<th>Index</th>
<th>Haplotype</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CAGTGTCCGTC</td>
<td>0.217</td>
</tr>
<tr>
<td>2</td>
<td>CAATGTCACTC</td>
<td>0.146</td>
</tr>
<tr>
<td>3</td>
<td>CAGCAGGAGTC</td>
<td>0.140</td>
</tr>
<tr>
<td>4</td>
<td>TGGCAGGACGT</td>
<td>0.096</td>
</tr>
<tr>
<td>5</td>
<td>TGGTGTCCGTC</td>
<td>0.090</td>
</tr>
<tr>
<td>6</td>
<td>TGGCAGGAGTC</td>
<td>0.085</td>
</tr>
<tr>
<td>7</td>
<td>CAGCAGGACGT</td>
<td>0.056</td>
</tr>
<tr>
<td>8</td>
<td>CGGCAGGAGCT</td>
<td>0.033</td>
</tr>
<tr>
<td>9</td>
<td>CAGTGTCACTC</td>
<td>0.031</td>
</tr>
<tr>
<td>10</td>
<td>CAGCAGGACTC</td>
<td>0.027</td>
</tr>
<tr>
<td>11</td>
<td>CGGTAGGACGT</td>
<td>0.010</td>
</tr>
<tr>
<td>12</td>
<td>CAGCAGGAGTT</td>
<td>0.009</td>
</tr>
<tr>
<td>13</td>
<td>CGGCAGGACGT</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Haplotypes shown in bold also occur in the full-sib family.
Table 2. *EniCOBL4A* associations with cellulose and kraft pulp yield (KPY) in a full-sib family of *E. nitens* (n = 275)

<table>
<thead>
<tr>
<th>Model</th>
<th>Genotype</th>
<th>n</th>
<th>Cellulose Mean (s.e.)</th>
<th>p-value</th>
<th>KPY Mean (s.e.)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codominant</td>
<td>GG</td>
<td>80</td>
<td>38.45 (0.11)</td>
<td>0.0006</td>
<td>50.83 (0.13)</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>138</td>
<td>38.94 (0.07)</td>
<td></td>
<td>51.32 (0.10)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>57</td>
<td>38.84 (0.12)</td>
<td></td>
<td>51.18 (0.14)</td>
<td></td>
</tr>
<tr>
<td>Dominant</td>
<td>GG</td>
<td>80</td>
<td>38.45 (0.11)</td>
<td>0.0002</td>
<td>50.83 (0.13)</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>AG-AA</td>
<td>195</td>
<td>38.91 (0.06)</td>
<td></td>
<td>51.28 (0.08)</td>
<td></td>
</tr>
<tr>
<td>Recessive</td>
<td>GG-AG</td>
<td>218</td>
<td>38.76 (0.06)</td>
<td>0.57</td>
<td>51.14 (0.08)</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>57</td>
<td>38.84 (0.12)</td>
<td></td>
<td>51.18 (0.14)</td>
<td></td>
</tr>
<tr>
<td>Overdominant</td>
<td>AA-GG</td>
<td>137</td>
<td>38.61 (0.08)</td>
<td>0.0032</td>
<td>50.98 (0.10)</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>138</td>
<td>38.94 (0.07)</td>
<td></td>
<td>51.32 (0.10)</td>
<td></td>
</tr>
<tr>
<td>Log-additive</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.0058</td>
<td>-</td>
<td>0.041</td>
</tr>
</tbody>
</table>

Significant *p*-values using ANOVA are highlighted in bold. *P*-value of the marker under codominant model is significant after correcting for chromosome-wide multiple testing.
Table 3. *EniCOBL4A* associations with cellulose and kraft pulp yield (KPY) in an association population of *E. nitens*. based on single marker analysis and haplotype analysis.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Cellulose</th>
<th>KPY</th>
<th>Haplotype</th>
<th>Cellulose</th>
<th>KPY</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>PrSNP-SNP2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PrSNP</td>
<td>0.79</td>
<td>0.94</td>
<td></td>
<td>0.20</td>
<td>0.08</td>
</tr>
<tr>
<td>SNP1</td>
<td>0.38</td>
<td>0.44</td>
<td>SNP1-SNP3</td>
<td>0.38</td>
<td>0.14</td>
</tr>
<tr>
<td>SNP2</td>
<td>0.83</td>
<td>1.00</td>
<td>SNP2-SNP4</td>
<td>0.72</td>
<td>0.35</td>
</tr>
<tr>
<td>SNP3</td>
<td>1.00</td>
<td>1.00</td>
<td>SNP3-SNP5</td>
<td>0.59</td>
<td>0.81</td>
</tr>
<tr>
<td>SNP4</td>
<td>1.00</td>
<td>0.99</td>
<td>SNP4-SNP6</td>
<td>0.58</td>
<td>0.59</td>
</tr>
<tr>
<td>SNP5</td>
<td>1.00</td>
<td>1.00</td>
<td>SNP5-SNP7</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td>SNP6</td>
<td>1.00</td>
<td>0.98</td>
<td>SNP6-SNP8</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>SNP7</td>
<td>0.014</td>
<td>0.010</td>
<td>SNP7-SNP9</td>
<td>0.035</td>
<td>0.046</td>
</tr>
<tr>
<td>SNP8</td>
<td>0.98</td>
<td>0.78</td>
<td>SNP8-SNP10</td>
<td>0.73</td>
<td>0.59</td>
</tr>
<tr>
<td>SNP9</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNP10</td>
<td>0.65</td>
<td>0.86</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p*-values using ANOVA are experiment-wise values adjusted for multiple testing based on permutation tests. *b*-values are based on permutation tests.
Table 4. SNP7 associations with cellulose and kraft pulp yield (KPY) in an association population of *E. nitens* (n = 416)

<table>
<thead>
<tr>
<th>Model</th>
<th>Genotype</th>
<th>n</th>
<th>Cellulose Mean (s.e.)</th>
<th>KPY Mean (s.e.)</th>
<th>p-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codominant</td>
<td>AA</td>
<td>187</td>
<td>40.85 (0.08)</td>
<td>52.60 (0.08)</td>
<td>0.023</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>188</td>
<td>40.58 (0.08)</td>
<td>52.29 (0.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>41</td>
<td>40.44 (0.19)</td>
<td>52.31 (0.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dominant</td>
<td>AA</td>
<td>187</td>
<td>40.85 (0.08)</td>
<td>52.60 (0.08)</td>
<td>0.0079</td>
<td>0.0037</td>
</tr>
<tr>
<td></td>
<td>AG-GG</td>
<td>229</td>
<td>40.55 (0.07)</td>
<td>52.29 (0.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recessive</td>
<td>AA-AG</td>
<td>375</td>
<td>40.71 (0.06)</td>
<td>52.44 (0.05)</td>
<td>0.14</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>41</td>
<td>40.44 (0.19)</td>
<td>52.31 (0.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overdominant</td>
<td>AA-GG</td>
<td>228</td>
<td>40.78 (0.08)</td>
<td>52.54 (0.07)</td>
<td>0.076</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>AG</td>
<td>188</td>
<td>40.58 (0.08)</td>
<td>52.29 (0.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-additive</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.007</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Significant *p*-values using ANOVA are highlighted in bold.
Figure Legends:

FIGURE 1. Phylogenetic tree of COBRA-like homologs.
Phylogenetic tree of the predicted COBRA-like (COBL) proteins from Arabidopsis, AtCOBRA (AT5G60920), AtCOBL1 (AT3G02210), AtCOBL2 (AT3G29810), AtCOBL4 (AT5G15630) AtCOBL5 (AT5G60950), AtCOBL6 (AT1G09790), AtCOBL7 (AT4G16120), AtCOBL9 (AT5G49270), AtCOBL8 (AT3G16860), AtCOBL10 (AT3G20580), AtCOBL11 (AT4G27110); Eucalyptus nitens, EniCOBL4A (FJ213604), EniCOBL4B (EW688390); Zea mays, Zmbk2 (EF078702); Oryza sativa, OsBC1 (AY328910).
Scale bar represents genetic distance. Node numbers indicate bootstrap support values.

FIGURE 2. EniCOBL4A gene structure and pair-wise linkage disequilibrium ($R^2$).
Numbers above the gene indicate the positions of SNPs in base pairs relative to the predicted transcription start site. Green indicates exons, white indicates introns and untranslated regions.

FIGURE 3. Chromosomal inheritance of EniCOBL4A haplotypes in Eucalyptus nitens full-sib family
Haplotypes determined from genotypes of PrSNP and SNP1 to 10 are shown. SNP7 is shown in bold and underlined. The letter M on haplotype 3 indicates methylated allele. Numbers beside the bars indicate parental haplotypes. Haplotype 2 and 4 are the same. Haplotype 4 is shown as 4^2. NN – northern New South Wales, MA – Macalister, SN – southern New South Wales, TO - Tooronga.

FIGURE 4. SNP7 genotype associations and position of cellulose QTL in full-sib family
(A). SNP7 associations with cellulose using four genotype classes. Error bars are standard error of mean (B). E. nitens linkage group (LG) 10 with QTL for cellulose. Reciprocals of $P$-values are shown on the right side of LG10.

FIGURE 5. Allelic expression imbalance in Eucalyptus nitens association population.
(A) Allelic expression imbalance in 16 unrelated SNP7 heterozygous trees determined using SNP7. Haplotypes were estimated and assigned to each tree using PHASE software with genotype information from SNPs 2 to 10 and are shown in brackets under each genotype class. Genotype of tree id 16 could not be determined. (B) Allelic expression imbalance in 9 SNP7 heterozygous trees belonging to genotype class 1 (Fig. 4A) and a single tree homozygous at SNP7 but otherwise the same genotype as genotype class 1. Allelic expression in all trees was measured using SNP4.
Error bars are standard error of mean. Mean value is based on four replicates. Dashed line indicates average allelic expression of gDNA controls. Allelic expression of genotype class 1 is significantly lower ($p < 0.001$) than gDNA controls using Kruskal-Wallis analysis with Dunn’s post test.

FIGURE 6. Electrophoretic mobility shift assay (EMSA) using probes specific to SNP7 alleles

(A). EMSA. Biotin labelled probes are incubated with (+) and without (-) nuclear extracts (NE) and the specificity of the reaction tested with 100 fold excessive molar concentration of the unlabelled competitive probe (com).

(B). EMSA with different probe concentrations. Differential binding of the nuclear extracts to the two alleles was repeated twice with different concentrations (750 fmol in lane 4 and 500 fmol in lane 6) of biotin labelled probes.

FIGURE 7. Allelic expression imbalance in the *Eucalyptus nitens* full-sib family.

Allelic expression was measured using cDNA from 28 trees heterozygous for SNP7. Heterozygous trees are made up of two different genotypic classes. Error bars are standard error of mean. Mean value is based on four replicates. Dashed line indicates average allelic expression of gDNA controls. Allelic expression of genotype class 23 is significantly higher ($p < 0.0001$) than gDNA controls using Kruskal-Wallis analysis with Dunn’s post test.
$R^2 > 0.75$

$R^2 < 0.45$
$$P = 2.00 \times 10^{-4}$$
$$n = 275$$
Genotype class 1
(GTGTCG/GCAGGA/GTC)

Genotype class 2
(GTGTCG/GCAGGA/GTC)

Genotype class 3
(ATTGC/GTC/GTGTCG/GTC)

 Allelic expression Imbalance (A/G)

Tree ID

Genotype class 1A
(GTGTCG/GCAGGA/GTC)

Genotype class 1A
(GTGTCG/GCAGGA/GTC)

Tree ID

Allelic expression Imbalance (A/G)
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>-</th>
<th>-</th>
<th>+</th>
<th>-</th>
<th>-</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>NE</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>SNP7 probe</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Allele A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Allele G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>